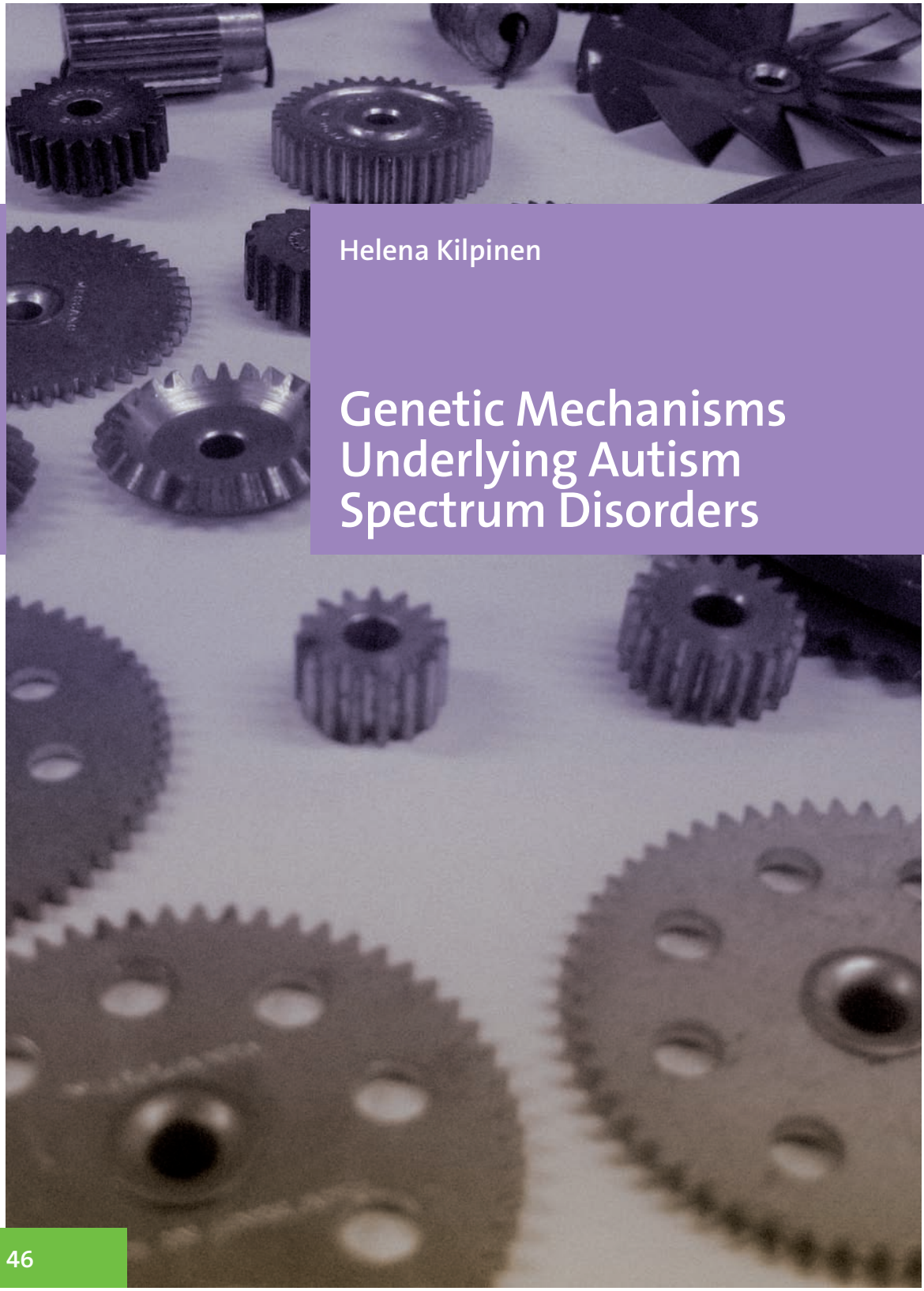
NATIONAL INSTITUTE
FOR HEALTH AND WELFARE



RESEARCH

Helena Kilpinen

Genetic Mechanisms Underlying Autism Spectrum Disorders

Helena Kilpinen

GENETIC MECHANISMS UNDERLYING AUTISM SPECTRUM DISORDERS

A C A D E M I C D I S S E R T A T I O N

To be presented with the permission of the Faculty of Medicine, University of Helsinki, for public examination in the Large Lecture Hall, Haartman Institute, on December 15th 2010, at 12 noon.

National Institute for Health and Welfare,

Institute for Molecular Medicine Finland (FIMM),

Department of Medical Genetics, University of Helsinki,

Research Program of Molecular Neurology, University of Helsinki,

and

Wellcome Trust Sanger Institute, Cambridge, United Kingdom

Helsinki 2010

Cover graphic: Jeffrey Barrett 2010

ISBN 978-952-245-376-1 (printed)
ISBN 978-952-245-377-8 (pdf)
ISSN 1798-0054 (printed)
ISSN 1798-0062 (pdf)

Printed by Helsinki University Print
Helsinki, Finland 2010

Supervised by

Academician of Science, Professor Leena Peltonen-Palotie
Wellcome Trust Sanger Institute
Cambridge, United Kingdom

Institute for Molecular Medicine (FIMM),
University of Helsinki
Helsinki, Finland

and

Docent Iiris Hovatta
Research Program of Molecular Neurology
University of Helsinki
Helsinki, Finland

Department of Mental Health and Substance Abuse Services
National Institute for Health and Welfare
Helsinki, Finland

Reviewed by

Docent Minna Männikkö
Department of Medical Biochemistry and Molecular Biology
University of Oulu
Oulu, Finland

and

Professor Juha Partanen
Department of Biosciences
University of Helsinki
Helsinki, Finland

Opponent

Professor Mark McCarthy
Oxford Centre for Diabetes, Endocrinology and Metabolism
Oxford University
Oxford, United Kingdom

"Somewhere ages and ages hence:
Two roads diverged in a wood, and I-
I took the one less traveled by,
And that has made all the difference. "
-Robert Frost, 1916

To Leena

ABSTRACT

Helena Kilpinen, Genetic Mechanisms Underlying Autism Spectrum Disorders. National Institute for Health and Welfare (THL), Research 46/2010, 199 pages. Helsinki, Finland 2010.
ISBN 978-952-245-376-1 (printed), ISBN 978-952-245-377-8 (pdf)

Autism is a severe childhood-onset developmental disorder characterized by deficits in reciprocal social interaction, verbal and non-verbal communication, and dependence on routines and rituals. It belongs to a spectrum of disorders (autism spectrum disorders, ASDs) which share core symptoms but show considerable variation in severity. The whole spectrum affects approximately 0.6-0.7% of children worldwide, inducing a substantial public health burden and causing suffering to the affected families. Despite having a very high heritability, ASDs have shown exceptional genetic heterogeneity, which has complicated the identification of risk variants and left the etiology largely unknown. However, recent studies have identified an increasing number of rare genetic causes responsible for the phenotype in individual families, suggesting that rare, family-specific events contribute significantly to the genetic basis of ASDs. In this study, we have studied the genetic basis of autism spectrum disorders in Finnish families both genome-wide and from the perspective of a single candidate gene.

First, we focused on the *DISC1* (*Disrupted-in-schizophrenia-1*) gene on chromosome 1q42, which is one of the most studied candidate genes in neuropsychiatric genetics. It was originally described in a large Scottish pedigree with a balanced translocation disrupting the gene co-segregating with a variety of psychiatric conditions. Subsequent linkage, association, and functional studies have implicated *DISC1* in the regulation of multiple aspects of embryonic and adult neurogenesis, and established its pathogenic role in multiple neuropsychiatric disorders across populations. We studied, for the first time, the role of *DISC1* in ASDs, and identified association with markers and haplotypes previously associated with psychiatric phenotypes in the Finnish population. We identified four polymorphic micro-RNA target sites in the 3'UTR of *DISC1*, and showed that hsa-miR-559 regulates *DISC1* expression *in vitro* in an allele-specific manner.

Second, we describe an extended autism pedigree with genealogical roots in Central Finland reaching back to the 17th century. To take advantage of the beneficial characteristics of population isolates to gene mapping and reduced genetic heterogeneity observed in distantly related individuals, we performed a microsatellite-based genome-wide screen for linkage and linkage disequilibrium in this pedigree. We identified a putative autism susceptibility locus on chromosome 19p13.3 and obtained further support for previously reported loci at 1q23 and 15q11-q13. Most promising candidates were two *transducin-like enhancer of split (E(sp1) homolog, Drosophila)* genes (*TLE2* and *TLE6*), clustered on 19p13, and *ATPase, Na+/K+ transporting, alpha 2 polypeptide* (*ATP1A2*) on 1q23.

To follow-up this study, we extended our study sample from the same sub-isolate and initiated a genome-wide analysis of homozygosity and allelic sharing using high-density SNP markers. We also analyzed global gene expression in lymphocytes of these individuals, and performed pathway analysis of SNP and gene expression data to comprehensively investigate the genetic cause of ASDs in these individuals and to gain information of the underlying biological processes. We identified a small number of haplotypes shared by different subsets of the genealogically connected cases, along with convergent biological pathways from SNP and gene expression data, which highlighted axon guidance molecules in the pathogenesis of ASDs.

In conclusion, the results obtained in this thesis show that multiple distinct genetic variants are responsible for the ASD phenotype even within single pedigrees from an isolated population. We suggest that targeted resequencing of shared haplotypes, linkage regions, and other susceptibility loci is essential to identify the causal variants and to understand the underlying mutational spectrum in autism spectrum disoders. We also report a possible micro-RNA mediated regulatory mechanism, which might partially explain the wide-range neurobiological effects of the *DISC1* gene.

Keywords: Autism spectrum disorders, Asperger syndrome, isolated population, *DISC1*, miRNA, linkage analysis, gene expression, GWAS, pathway analysis

TIIVISTELMÄ

Helena Kilpinen, Genetic Mechanisms Underlying Autism Spectrum Disorders [Autismikirjon häiriöiden geneettiset mekanismit]. Terveyden ja hyvinvoinnin laitos, (THL), Tutkimus 46/2010, 199 sivua. Helsinki 2010.
ISBN 978-952-245-376-1 (printed), ISBN 978-952-245-377-8 (pdf)

Autismi on vakava lapsuusiän kehityksellinen häiriö, jonka tyypillisiä oireita ovat ongelmat vastavuoroisessa sosiaalisessa vuorovaikutuksessa, sanallisessa ja sanattomassa kommunikaatiossa, sekä riippuvuus erilaisista rutiineista ja rituaaleista. Se kuuluu laajempaan autismikirjon häiriöiden ryhmään, jotka kaikki ovat perusoireiltaan samankaltaisia, mutta joiden vakavuusaste vaihtelee huomattavasti. Autismikirjon häiriöitä tavataan maailmanlaajuisesti noin 0,6-0,7 %:lla lapsista, ja ne aiheuttavat inhimillisen kärsimyksen lisäksi merkittäviä kuluja terveydenhuollolle. Vaikka autismikirjon häiriöt ovat vahvasti perinnöllisiä, niiden geneettinen tausta on osoittautunut erittäin heterogeeniseksi, mikä on vaikeuttanut altistavien geneettisten tekijöiden tunnistamista ja perimmäisten syiden selvittämistä. Viimeaikaisissa tutkimuksissa on kuitenkin löydetty useita harvinaisia geneettisiä tekijöitä, jotka riittävät aiheuttamaan häiriön yksittäisissä perheissä. Näiden löydösten perusteella on arvioitu, että harvinaiset, perhespesifiset geneettiset muutokset selittävät merkittävän osan autismin periytyvyydestä. Tässä väitöskirjassa olemme tutkineet autismikirjon häiriöiden geneettistä taustaa suomalaisissa perheissä sekä koko genomin laajuisesti, että yksittäisen ehdokasgeenin näkökulmasta.

Ensimmäinen tutkimuskohteemme oli *DISC1*-geeni (*Disrupted-in-schizophrenia-1*), joka sijaitsee kromosomissa 1q42 ja joka on yksi tunnetuimmista geeneistä psykiatrisen genetiikan alalla. Se kuvattiin alunperin suuressa skotlantilaisessa suvussa, jossa balansoitu translokaatio periytyi yhdessä erilaisten psykiatristen sairauksien kanssa. Tätä seuranneet kytkentä-, assosiaatio-, sekä toiminnalliset tutkimukset ovat osoittaneet, että *DISC1* säätelee lukuisia hermostokehityksellisiä tomintoja sekä alkionkehityksen aikana että aikuisiällä, ja useissa eri populaatioissa on havaittu, että muutokset *DISC1*:n toiminnassa voivat altistaa monille neuropsykiatrisille sairauksille. Tässä väitöskirjassa tutkittiin ensimmäisen kerran *DISC1*-geenin merkitystä autismikirjon häiriöissä. Havaitsimme, että samat geenimerkit ja haplotyypit, joiden on aikaisemmin osoitettu assosioituvan muihin psykiatrisiin fenotyyppeihin suomalaisväestössä, assosioituivat myös autismikirjon häiriöihin. Havaitsimme lisäksi, että geenin 3'UTR aluella sijaitsee neljä polymorfista mikro-RNA:n (miRNA) sitoutumiskohtaa, ja osoitimme, että hsa-miR-559 säätelee *DISC1*:n ilmentymistä alleelispesifisesti *in vitro*. Saamiemme tulosten perusteella *DISC1*:n geneettiset variantit altistavat autismikirjon häiriöille ja vaikuttavat geenin miRNA-välitteiseen säätelyyn.

Toiseksi, olemme kuvanneet yksittäisen suuren autismisuvun, jonka on havaittu polveutuvan yhteisistä keskisuomalaisista esivanhemmista 1600-luvulla. Hyödyntääksemme populaatioisolaattien geenikartoitukselle tarjoamia mahdollisuuksia, teimme

genominlaajuisen kytkentä- ja kytkentäepätasapainoanalyysin mikrosatelliittimarkkereilla tässä suvussa, jonka kaukaisen sukulaisuuden voidaan olettaa vähentävän geneettistä heterogeenisyyttä. Tunnistimme mahdollisen autismille altistavan lokuksen kromosomissa 19p13.3, ja havaitsimme kaksi jo aiemmin autismiin yhdistettyä lokusta kromosomeissa 1q23 ja 15q11-q13. Lupaavimmat ehdokasgeenit olivat *TLE2* ja *TLE6 (transducin-like enhancer of split [E(sp1) homolog, Drosophila])*, jotka sijaitsevat 19p13-lokuksessa, sekä *ATP1A2* (*ATPase, Na+/K+ transporting, alpha 2 polypeptide)* kromosomissa 1q23. Tulokset osoittivat, että autismin geneettinen tausta yksittäisessäkin suvussa on monitekijäinen, ja että altistavia geneettisiä muutoksia on tässä suvussa ainakin kolmella eri kromosomialueella.

Selvittääksemme autismikirjon häiriöiden geneettistä ja biologista taustaa mahdollisimman kattavasti ja tutkiaksemme edellä saatuja tuloksia tarkemmin, laajensimme käyttämäämme keskisuomalaistaustaista aineistoa uusilla perheillä, joilla on samankaltaiset genealogiset juuret. Teimme genominlaajuisen analyysin autismia sairastavien henkilöiden homotsygoottisista genomisista alueista sekä alleelisesta jakamisesta käyttämällä tiheästi sijoittuvia SNP-markkereita. Analysoimme lisäksi genominlaajuista geenien ekspressiota saman aineiston henkilöiden valkosoluista, ja kartoitimme alustavasti mahdollisia altistavia biologisia reaktioreittejä käyttämällä sekä SNP- että geenien ekspressiotuloksia. Tunnistimme pienen määrän haplotyyppejä, jotka osa sukupuun autismia sairastavista henkilöistä jakoi keskenään, sekä yhteneviä reaktioreittejä SNP- ja ekspressiodatasta, joiden perusteella aksoniohjaukseen osallistuvat molekyylit näyttäisivät liittyvän autismikirjon häiriöiden patogeneesiin.

Yhteenvetona, tässä väitöskirjatyössä saadut tulokset osoittavat, että useat erilliset geneettiset muutokset johtavat autismikirjon häiriöihin, jopa populaatioisolaateissa ja yksittäisissä perheissä. Näiden tulosten perusteella voidaan sanoa, että kytkentäalueiden, jaettujen haplotyyppien, sekä muiden altistavien genomisten alueiden sekvensointi on välttämätöntä, jos halutaan tunnistaa varsinaiset mutaatiot ja variantit, jotka aiheuttavat autismikirjon häiriöitä. Tunnistimme myös mahdollisen mikroRNA-välitteisen säätelymekanismin, joka saattaa osittain selittää *DISC1*-geenin laaja-alaisia neurobiologisia vaikutuksia.

Avainsanat: autismikirjon häiriöt, Aspergerin oireyhtymä, populaatioisolaatti, *DISC1*, miRNA, kytkentäanalyysi, geenien ekspressio, GWAS, reaktioreittianalyysi

CONTENTS

ABBREVIATIONS

ACC	Autism Case Control (Cohort)
ADI-R	Autism Diagnostic Interview - Revised
ADOS	Autism Diagnostic Observation Schedule
AGP	Autism Genome Project
AGRE	Autism Genetic Resource Exchange
AS	Asperger syndrome
ASD	autism spectrum disorder
ASDI	Asperger Syndrome Diagnostic Interview
ASP	affected sib pair
ASSQ	Asperger Syndrome Screening Questionnaire
bp	base pair
BPD	bipolar disorder
cAMP	cyclic adenosine monophosphate
CARS	Childhood Autism Rating Scale
CD	Crohn's disease
CDCV	common disease common variant
cDNA	complementary DNA
CF	Central Finland
CGH	comparative genomic hybridization
ChIP	chromatin immunoprecipitation
CNV	copy number variation
cM	centiMorgan
CNS	congenital nephrotic syndrome
CNV	copy number variation
dNTP	deoxynucleotide
ddNTP	dideoxynucleotide
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, 4th Edition
DZ	dizygotic
eQTL	expression quantitative trait locus
EDTA	ethylenediaminetetraacetic acid
ECARUCA	European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations
fMRI	functional magnetic resonance imaging
FRAXA	fragile X syndrome
GABA	gamma-aminobutyric acid
GSEA	gene set enrichment analysis
GWAS	genome-wide association study
hME	homogeneous MassEXTEND
HWE	Hardy-Weinberg equilibrium
IBD	identical by descent

IBS	identical by state
ID	intellectual disability
ICD-10	International Classification of Diseases, 10th Revision
IMGSAC	International Molecular Genetic Study of Autism Consortium
IQ	intelligence quotient
kb	kilobase
LC	liability class
LCL	lymphoblastoid cell line
LD	linkage disequilibrium
LOD	logarithm of odds
MAF	minor allele frequency
MALDI-TOF	matrix-assisted laser desorption/ionization time-of-flight
Mb	megabase
MIM	Mendelian Inheritance in Man
miRISC	miRNA-induced silencing complex
miRNA	micro-RNA
MLS	maximum LOD score
mRNA	messenger RNA
MZ	monozygotic
NPL	non-parametric LOD
PCR	polymerase chain reaction
PDD	pervasive developmental disorder
PDD-NOS	pervasive developmental disorder not otherwise specified
QC	quality control
QTL	quantitative trait locus
RMA	robust multiarray average
ROH	region of homozygosity
SCZ	schizophrenia
SNP	single nucleotide polymorphism
TDT	transmission disequilibrium test
TF	transcription factor
TS	Tourette syndrome
TSC	tuberous sclerosis
UTR	untranslated region
WHO	World Health Organization
WTCCC	Wellcome Trust Case Control Consortium
Zmax	maximum LOD score
θ	recombination fraction

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original articles referred to in the text by their Roman numerals. In addition, some previously unpublished data are presented.

I **Kilpinen H**, Ylisaukko-oja T, Hennah W, Palo OM, Varilo T, Vanhala R, Nieminen-von Wendt T, von Wendt L, Paunio T, Peltonen L (2008) Association of DISC1 with autism and Asperger syndrome. *Molecular Psychiatry*, 13:187-96.

II **Kilpinen H**, Ylisaukko-oja T, Rehnström K, Gaál E, Turunen JA, Kempas E, von Wendt L, Varilo T, Peltonen L (2009) Linkage and linkage disequilibrium scan for autism loci in an extended pedigree from Finland. *Human Molecular Genetics*, 18:2912-21

III **Kilpinen H***, Rehnström K*, Rossi M, Saharinen J, Jakkula E, Greco D, Ylisaukko-oja T, Ripatti S, Daly MJ, Purcell S, Auvinen P, Varilo T, von Wendt L, Barrett JC, Palotie A, Hovatta I, Peltonen L (2010) Genetic architecture of extended autism pedigrees from a population isolate point to genes involved in axon guidance. *Submitted.*

*These authors contributed equally to this work.

Publication III has appeared previously in the doctoral thesis of Karola Rehnström (2009).

1 INTRODUCTION

Childhood autism is the hallmark diagnosis of a spectrum of similar neurodevelopmental conditions, known as autism spectrum disorders (ASD). Autism was first described in 1943 by Leo Kanner, who reported children with intellectual disability and severe social isolation which was not explained by the developmental level of the children (Kanner 1943). It is characterized by various impairments in behavioral, communicational, and social skills, such as dependence on routines and rituals, and absence or delay of spoken language. Asperger syndrome (AS), part of the same spectrum, was described shortly after autism in 1944 by Hans Asperger, who described patients with "autistic psychopathy" but normal intellectual abilities (Asperger 1944). Altogether, ASDs affect 0.6-0.7 % worldwide and are amongst the most heritable of neuropsychiatric disorders. Autism has traditionally been considered a common complex disorder, and studies investigating its genetic component have been conducted since the 1980's.

Until recently, very little was known about the genetic basis of autism spectrum disorders, much like many other psychiatric and neurological disorders. Based on information from linkage analysis and candidate gene studies it was thought that most of the predisposing genetic variants would be common in the population, and an unspecified combination of both genetic and environmental factors would be required to cross the critical threshold of liability. However, the identification of genes and genetic variants predisposing to ASDs has been significantly complicated by their exceptional heterogeneity. Phenotypically, significant differences are seen across the core clinical features, and some individuals with ASDs have co-occurring medical conditions with known etiologies, such as mendelian genetic defects. Additionally, the extent of genetic heterogeneity in ASDs has only recently been fully revealed, suggesting that many of the predisposing genetic factors may be unique to specific families. Reflecting this, it has been suggested that instead of treating autism as a single diagnosis, it would be more appropriate to think about "the autisms" (Geschwind and Levitt 2007).

It is interesting to note in this context that genetic findings that overlap among different neuropsychiatric disorders have been increasingly discovered. For example, association of genetic variants in the *DISC1* gene (*Disrupted-in-schizophrenia-1,* originally associated to schizophrenia) to a broad range of phenotypes, and the discovery of same, rare DNA copy number variants in schizophrenia and ASDs, has forced us to reconsider the usefulness of traditional diagnostic classifications. These observations also underscore the likelihood that the underlying biological processes among these distinct clinical phenotypes will be at

least partially convergent. The idea is further supported by this study, where we found that genetic variants of the *DISC1* gene associate also to ASDs.

This project was started during the early years of the "post-genomic" era and the ongoing refinement of the clinical (and genetic) relationship among "the autisms" and other neuropsychiatric traits. Whereas previously only targeted research questions about the genetic etiology of complex disease could be addressed, it is now possible to use approaches which query the entire genome of an individual. In this thesis, we have taken advantage of multiple layers of genome-wide data and the properties of the isolated Finnish population to explore the genetic and biological cause of ASDs in a set of genealogically connected individuals. As a result, a complex picture of multiple layers of genome biology has slowly started to emerge and provide cues of the mechanisms underlying various genetic traits and disorders.

2 REVIEW OF THE LITERATURE

2.1 Variation in the genome and complex diseases

2.1.1 Genetic variation

All human genomes are approximately 99.9% identical, and the remaining 0.1% of variation between individuals together with environmental factors is what makes people different from each other. Genetic variation comes in multiple forms, and the variable sites of the genome can be used as genetic markers to study phenotypic differences between individuals and populations. Of particular interest is the study of disease genetics, which aims to identify, or map, genetic differences in the genome, which determine whether an individual is susceptible to a disease or not. It has been known for long that certain phenotypic traits and diseases run in families, and can be transmitted from parents to offspring. Although familial is not a synonym for heritable, as shared environment alone can be sufficient to produce a shared phenotype, familial transmission of some traits was eventually found to be mediated by specific genetic differences which children inherited from their parents.

Genetic variation among humans arises mainly through two distinct mechanisms. Mutations are introduced to the genome randomly, with an average frequency of ~ 2.2×10^{-9} per base pair per year (Kumar and Subramanian 2002), although some regions of the genome are more vulnerable to mutations than others. Mutations are classified as point mutations, insertions, or deletions, based on their effect on the gene structure. They can also be classified as loss-of-function, gain-of-function, or neutral mutations, based on the effect they have on the function of the gene in question. Mutations are introduced to the genome either by external factors, such as mutagenic radiation or chemicals, or by internal factors such as occasional random errors in DNA replication during cell division. Another biological process that creates variation among humans is recombination. Recombination refers to the process in which a DNA molecule (or sometimes RNA) is broken and then joined to a different one. Homologous recombination occurs between similar molecules of DNA, usually during mitosis, whereas recombination during meiosis facilitates chromosomal crossover which creates novel, unique combinations of the parental genomes which are then transmitted to the offspring.

Genetic variation in humans can roughly be divided into structural variation and sequence variation. While structural variation is a broad term encompassing many different types of rearrangements affecting the physical structure of the chromosome, such as deletions and duplications of regions of varying length,

sequence variation occurs only on the DNA sequence-level, i.e. on the level of single base pairs comprising the "genetic code". In the early days, available cytogenetic methods only allowed the detection of large structural variants, such as deletions of entire chromosome arms, whereas new technologies have made it possible to detect variation in DNA copy number (copy number variants, CNVs) at much higher resolution (Feuk *et al.* 2006). This has lead to studies showing that small deletions, duplications, and other types of structural variation are abundant in the human genome and their effect is often benign (Iafrate *et al.* 2004, Sebat *et al.* 2004, Conrad *et al.* 2006, Redon *et al.* 2006, Conrad *et al.* 2010). Similarly, when DNA sequencing became common practice, it became clear that different kinds of sequence polymorphisms, such as single nucleotide polymorphisms (SNPs), are common in the genome, and contribute to the normal variation between individuals.

2.1.2 Genetic mapping

To determine whether a trait has a genetic component, the degree of its heritability needs to be determined. Heritability is defined as the proportion of the phenotypic variance which is explained by genetic factors. Estimates are by rule produced using family, twin, and adoption studies, for example by comparing concordance rates in monozygotic (MZ) and dizygotic (DZ) twins which share 100% or 50% of their genome, respectively. Once the genetic basis of a trait has been established, genetic mapping can be applied to identify genomic regions, genes, and specific genetic variants in the genome that cause or contribute to the phenotype. Genetic mapping takes use of naturally occurring genetic markers and does not require prior knowledge of the pathogenesis of the disease or other trait in question. Instead, the aim is to look for regions in the genome that are genetically or statistically linked with the phenotype. Genetic mapping is facilitated by strong linkage disequilibrium (LD) (see Section 2.1.4) between the marker allele and the causative variant, as well as high penetrance and large effect size of the causal variant. If a heritable disorder is caused by defects in a single gene, they are referred to as monogenic, or mendelian, disorders, and their inheritance follows the general laws of inheritance, first discovered by Gregor Mendel in the 19th century. If instead the disorder is caused by the combined effect of multiple different genes or other genetic factors, it is referred to as a polygenic, or a complex disorder. Different methods for genetic mapping are described in Section 2.1.5.

A genetic marker is a variable, or polymorphic, segment of DNA with an identifiable physical location on a chromosome whose inheritance in pedigrees can be traced. To be suitable for genetic mapping studies a marker has to be sufficiently polymorphic. The most commonly used markers include SNPs and microsatellite markers (also called short tandem repeats). SNPs are mostly di-allelic and the most abundant type of variation in the genome. Based on the Phase II of the International

fr Project (see Section 2.1.4), the total number of common SNPs in the human genome is at lest ~9-10 million (Frazer *et al.* 2007). Microsatellies are multiallelic markers consisting of tens of copies of di-, tri- or tetranucleotide repeats, and therefore generally more informative for mapping. However, they are less abundant in the genome and suffer from a relatively high mutation rate of ~10^{-3} to 10^{-4} per locus per generation (Levinson and Gutman 1987, Weber and Wong 1993, Ellegren 2000).

2.1.3 Complex disease genetics

The methodology used in complex disease genetics was originally developed for the mapping of monogenic disorders. The traditional way to identify a causative variant or a mutation was to start with a genome-wide linkage analysis in a small set of affected families, fine-map the identified linkage region to narrow down the region of interest, and eventually identify the gene or variant at the locus by physical mapping or direct sequencing. As linkage analysis is best suited for the identification of rare, relatively high-penetrant variants, new methods recently evolved to analyze the common variation in the genome. An example of such methods is a genome-wide association study (GWAS) where allele frequencies of thousands of common SNPs are compared between cases and controls (see Section 2.1.5).

Two main hypotheses exist regarding the genetic background of common complex diseases. The first is that a small number of rare genetic variants (< 1% frequency in the general population), each with a large effect, cause the disease (Reich and Lander 2001). The second is that common traits are caused by a relatively large number of common genetic variants, each with a small effect on the phenotype (Lander 1996, Chakravarti 1999). This is called the "common disease common variant" (CDCV) hypothesis, an example of which is Alzheimer's disease, where a specific allele of the apolipoprotein E gene appears to be responsible for over 50% of cases (Corder *et al.* 1993). In reality, both types of variants are likely to contribute to the disease risk, as GWA studies have shown that common variation alone explains only a relatively small fraction of the overall genetic risk, and single common variants associated with a disease have at best modest effect sizes. In a recent study, the mean odds ratios were calculated for most rare and common alleles identified in complex diseases to date (Bodmer and Bonilla 2008). The mean odds ratio for rare alleles was 3.74 and for common alleles 1.36. As GWA studies tackle only common alleles (population frequency > 1%), the next wave of studies has already started to analyze rare variation, which will probably explains at least a part of the "missing heritability".

Overall, GWA studies in common complex diseases have been most successful in autoimmune diseases such as Crohn's disease, for which over 30 loci accounting for ~20% of the total heritability have been robustly identified (Barrett *et al.* 2008). This is a high percentage, since for most other common traits, such as height and serum lipid levels, the identified variants account for only ~5-10% of the heritability (Visscher 2008, Aulchenko *et al.* 2009). One of the most extensive GWAS efforts was the study by the Wellcome Trust Case Control Consortium (WTCCC), which analyzed 14 000 patient samples representing seven different common disorders and a set of 3000 shared controls (2007). The phenotypes included were Type I and Type II diabetes, Crohn's disease, rheumatoid arthritis, hypertension, coronary artery disease, and bipolar disorder. Genome-wide significant loci were identified for all diseases, except hypertension. The total number of GWA studies published to date is approaching 800 (www.genome.gov/GWAStudies). GWAS in neuropsychiatric phenotypes have been less successful, potentially because the overall complexity in brain-related phenotypes is much higher or because common variation plays a much smaller role in these phenotypes. Yet, GWA studies have completely transformed genetic mapping studies, and substantially increased, and altered, our understanding of the genetic background of complex disorders.

2.1.4 Extent of linkage disequilibrium

Linkage disequilibrium (LD) is defined as the non-random association of alleles at separate but linked loci on the same chromosome. In the genome, LD exists as haplotype blocks which vary in length and consist of regions with low recombination separated by regions of high recombination rates (referred to as recombination hotspots). As a general rule, the further apart two loci are on a chromosome the weaker is the LD between them, with different chromosomes showing complete independence of each other, that is, linkage equilibrium. However, the exact extent of LD at different loci is not predictable and thus, the LD structure of the human genome in different populations has been intensively studied, mostly because the extent and distribution of LD determines the number of markers required for a GWA study (Service *et al.* 2006). LD is normally a result of ancestral haplotypes being common in the population, and the blocks typically contain only a few common haplotypes (Daly *et al.* 2001, Reich *et al.* 2001, Gabriel *et al.* 2002, Phillips *et al.* 2003). LD is the basis of genetic association studies where genetic variants predisposing to the phenotype-of-interest can be identified by genotyping other markers which are in LD with the predisposing variant (see Section 2.1.5, Association-based methods).

A key effort in characterizing variation and LD in the human genome was undertaken by the International HapMap Project (www.hapmap.org) (2003, 2005), which set out to create a haplotype map of the entire genome. A haplotype map

illustrates the LD structure across all chromosomes and predicts which markers are inherited together. The HapMap project identified, validated, and genotyped common SNPs in different global populations, and provided a publicly available resource of SNPs to the scientific community. In the Phase I of the project, ~one million SNPs were genotyped in four populations (CEU – Central Europeans in Utah, representing a Caucasian population, CHB – Han Chinese from Bejing, JPT – Japanese from Tokyo, YRI – Yorubans from Nigeria) (The International HapMap Consortium 2005), whereas in Phase II, the number of SNPs was extended to 3.1 million (Frazer *et al.* 2007). In the currently ongoing Phase III, the number of analyzed individuals has been increased and an additional seven populations have been included.

The extent of LD in a population is determined by the effective population size and the time since founding, which affect the number of different haplotypes present in the population and thus, the number of recombinations between markers. LD is also strongly influenced by the number of founders and the expansion rate of the population (Service *et al.* 2006). In a study by Service and colleagues (2006), the magnitude and distribution of LD was compared in 11 isolated populations (see Section 2.1.6) and an outbred European-derived sample. The authors discovered that although the profiles of the LD maps were very similar in all of the populations studied, as discovered before (Tapper *et al.* 2003, De La Vega *et al.* 2005), the overall length of the maps, i.e. the extent of LD, varied. Interestingly, the most extensive LD was observed in a sub-isolate of Finland (named Kuusamo), and the authors suggested that GWA studies with Finnish samples could require as much as ~30% fewer markers than with samples from more outbred populations. However, this is merely a theoretical suggestion, since current GWA studies are mostly conducted using predesigned SNP platforms. Nevertheless, this observation proved that population isolates are beneficial for complex disease mapping.

2.1.5 Methods for genetic mapping in disease genetics

An overview of the most essential methods is given, with the emphasis on the methods used in this thesis.

Linkage analysis

Linkage is a genetic phenomenon in which different characters, such as a phenotype and a marker allele, co-segregate in a pedigree because their determinants lie close together on a particular chromosome and are not separated by recombination. The probability of linkage correlates with the genetic distance of two loci. The closer the loci are the smaller is the probability of meiotic recombination occurring between them. This recombination fraction (θ) ranges from 0.5 (complete linkage) to 0 (no

linkage), and the relationship between the recombination fraction and actual genetic distance is defined by a specific mathematical mapping function. The logarithm of odds (LOD) score represents the probability that two loci are inherited together because of linkage instead of chance (Morton 1955). However, recombination is not random and its rate is known to vary depending on for example sex and genomic location (Broman *et al.* 1998, Yu *et al.* 2001).

Linkage analysis is a family-based genetic analysis method to study whether two loci in the genome, i.e. the "disease locus" and the marker locus, are genetically linked together. The objective is to find a locus that is inherited together with the trait in question more often than it should by chance (Terwilliger and Ott 1994). Linkage analysis was one of the first methods developed for genetic mapping and it has traditionally been carried out using microsatellite markers, since these were among the first genetic markers to be discovered and, based on their properties, were well suited for genome-wide mapping. In a single-gene disease, linkage analysis is an extremely efficient way to pinpoint the risk locus, and further the gene, since the locus usually shows complete segregation with the trait. However, in genetically more complex diseases, multiple loci contribute to the disease risk and therefore, multiple linkage peaks are typically identified. Linkage analysis identifies only relatively large genomic regions, which is why a denser marker set needs to be subsequently genotyped at the regions in order to narrow down the region of interest.

Association-based methods

A. Candidate gene studies

While linkage is a relationship between genetic loci, association is simply a statistical observation that might have various non-genetic causes (see also B. Genome-wide association studies). Association studies operate on the population level, and can be regarded as very large linkage studies of unobserved, hypothetical pedigrees (Cardon and Bell 2001). A candidate gene association study has traditionally been the next step after a genome-wide linkage scan. Once a number of genomic regions have been identified by linkage, candidate genes are chosen from the linked region, finemapped typically with SNP markers, and analyzed for a correlation, i.e. a statistical association with the phenotype. The aim is to increase the density of markers in a targeted locus, and identify common variants that would associate to the phenotype and explain the observed linkage signal. If the finding is real, the associating marker allele presumably is in LD with the actual disease predisposing variant and the variant is flanked with a detectable haplotype (Figure 1). In rare cases, the associating allele might directly influence the risk of disease. Since linkage regions are typically large and contain tens or even hundreds of genes, it is usually not possible to analyze all of the genes. Thus, candidate genes are often

prioritized by prior biological information, picking the genes that seem most relevant to the phenotype in question. This is of course a biased approach, since the choice of genes is entirely dependent on existing data and the extent of knowledge of the person who picks the genes, and genes whose function is unknown or less well characterized, are often left out. Candidate gene association studies can also be performed on the basis of a specific biological hypothesis, instead of a positional hypothesis, i.e. a linkage analysis. In this case, a gene or a group of related genes is studied for association with the phenotype, because previous research information strongly supports the hypothesis that it would be involved in a particular disease based on its function. An example of this is an association study between 155 ion transport genes and migrane (Nyholt *et al.* 2008).

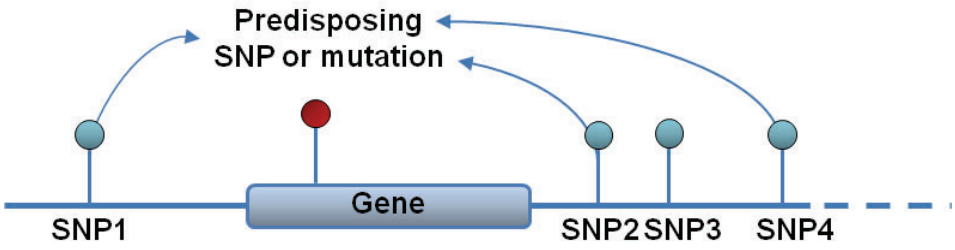


Figure 1. ***Illustration of linkage disequilibrium (LD) as the basis of genetic association studies.*** *A SNP or a mutation predisposing to a given phenotype (red) can be identified by genotyping other genetic markers (blue) which are in LD with the marker-of-interest (denoted by arrows). However, the degree of LD at a given locus is not always predictable or consistent with distance (denoted by SNP3 which is not in LD with the predisposing SNP despite its close proximity).*

B. Genome-wide association studies

Genome-wide association studies (GWAS) were spurred by decreasing genotype costs, which enabled the genotyping of hundreds of thousands of SNP markers across the genome in very large patient cohorts. Like linkage studies, GWAS is a hypothesis-free approach, but instead of looking for a truly genetic phenomenon like linkage, it is merely a statistical comparison of allele frequencies between two groups. The overall aim of a GWAS is to analyze as much of the common genetic variation in the genome as possible. Typically carried out in a case-control setting, GWA studies are designed to search for association of common genetic variants (minor allele frequency > 1%), since the SNP content of most of the commercial genotyping platforms reflect that of the International HapMap project, which has

identified and catalogued common SNPs and the LD structure across the genome in different populations. As the number of SNPs in HapMap has increased, so has the content of SNP chips, with the current platforms profiling around one million SNPs at a time. A method called imputation can additionally be used to increase the number of analyzed SNPs and to fill in missing genotypes. In imputation, allele frequency and LD information from other comparable genotyping studies can be used to estimate the most probable genotype at any given locus.

Due to the massive number of SNPs analyzed, the number of statistical tests in a GWAS is huge, which increases the burden of multiple testing. Thus, sufficient statistical power has become a key issue in GWA studies, leading to gigantic study samples which have already exceeded 100 000 in some phenotypes (Teslovich *et al.* 2010). Larger study samples are also achieved through meta-analysis, which is statistical technique to combine results from multiple different studies. The statistical power issue goes hand-in-hand with the effect size of the genetic variants, since it has now become clear that individual variants cause only a small increase to the risk in most phenotypes, and the smaller the effect, the more statistical power (i.e. larger study sample) is needed to detect it on a reliable, genome-wide significant level. A case-control analysis is the most common strategy, because individual affected cases are easier to obtain in large numbers than complete families, but some family-based studies have also been seen (for e.g. Weiss *et al.* 2009). The interest towards them and SNP-based linkage analyses has lately increased again, with the realization that common variants are not sufficient to explain the entire genetic component of common diseases.

Analyzing huge study samples can easily introduce bias into a GWA analysis. The most common sources of bias include confounding factors such as population stratification (see also Section 2.1.6), insufficient quality control of raw genotype data, or genotyping errors. Also, as mentioned, underpowered study samples can introduce spurious false-positive associations, which is why independent replication of results is an essential confirmatory step. The signal intensity data from genome-wide SNP genotyping platforms can also be used to analyze DNA copy number variation (CNV), i.e. a type of sub-microscopic structural variation of the genome.

Gene expression profiling and eQTLs

The process of cellular differentiation and development is primarily driven by the differential expression of genes. The relative abundance of specific transcripts in a cell at a given time is the key determinant for the function and developmental fate of the cell, and further, the whole organism. Differences in gene expression levels have been extensively studied to address the question of normal versus abnormal cellular state, and to characterize cellular differentiation processes by monitoring changes in gene expression profiles. In disease genetics, the aim of genome-wide, or global,

gene expression profiling is to determine the response of the studied tissue to the disease state and to identify those genes that are differentially expressed in the profiled tissue between disease cases and matched, healthy controls. However, one of the major challenges in gene expression studies has been the identification of an appropriate tissue or cell type for analysis, especially when studying phenotypes such as autism or other neuropsychiatric disorders, where the affected tissues are mostly unknown (Cole *et al.* 1999).

Especially in brain-related phenotypes, Epstein-Barr virus transformed B-lymphocytes (i.e. lymphoblastoid cell lines, LCLs) are the most common tissue used in gene expression profiling, although their use has often been criticized as being inappropriate. It has been argued that the immortalization of the cell line drastically changes the expression profile, introduces artifacts, and masks the possible true effects (Plagnol *et al.* 2008, Min *et al.* 2010). Also, blood-cell derived LCLs do not seem like the tissue-of-choice when looking for expression differences relevant to brain-related processes. However, there are also beneficial aspects of using LCLs. Firstly, and most importantly, they are easily accessible. Blood samples are relatively easy to obtain, as compared with other tissues, which enables larger sample sizes and better statistical power. Secondly, it has been argued that the transformation of the cells actually increases their comparability by removing effects of many environmental factors, such as smoking, which often affect the expression profile of native cells such as lymphocytes (Charlesworth *et al.* 2010). Thirdly, in many diseases, it is not at all clear which tissue would be more relevant to study. For example, expression profiling from brain tissue would require a very precise and informed hypothesis of the brain region most likely to be affected. This is especially true in autism, where the etiology remains poorly understood, and the causes and consequences of neurological and behavioral symptoms cannot be properly differentiated.

Recently, large scale studies have begun to address the effects of genetic variation on gene expression levels on a genome-wide level. By quantifying the transcript levels in a given tissue and correlating this information for example with SNP or microsatellite genotypes using linkage (Goring *et al.* 2007) or association analysis (Dixon *et al.* 2007, Stranger *et al.* 2007a, Stranger *et al.* 2007b, Kwan *et al.* 2008), quantitative trait loci (QTL) contributing to gene expression differences among populations, individuals, and tissues have been identified and are generally referred to as expression-QTLs (eQTL). This approach is especially useful in trying to determine how the variants identified in GWA studies might mediate disease susceptibility, most of which fall into gene deserts or lack apparent functionality. Accordingly, a recent study showed that current signals obtained from GWA studies are enriched for these regulatory variants (Nica *et al.* 2010).

Pathway analysis

A biological pathway is traditionally defined as a series of interconnected enzymatic steps linked by the production of intermediates that are then used in the following enzymatic step to produce a specific product. However, a biological pathway can also refer to various cellular signaling pathways where a single input signal, such a mechanical or a chemical stimulus, is converted into distinct cellular responses (i.e. signal transduction).

As the amount of various genome-wide datasets has quickly increased, biological pathway analysis has become a popular tool to address whether the most significant results are enriched for a particular group of genes with similar functions (i.e. belonging to a same pathway) more than would be expected by chance. To date, pathway analysis has mostly been applied to global gene expression datasets, but the last few years have seen an increasing number of studies, where similar methodology has been developed also for genome-wide association datasets (for e.g. Wang *et al.* 2007, Baranzini *et al.* 2009, Holmans *et al.* 2009, O'Dushlaine *et al.* 2009). In contrast to gene expression data where a single transcript can generally be used to represent a single protein product (i.e. a single component of a pathway), GWAS data is less straightforward in terms of pathway analysis. With multiple SNPs per gene, variable gene size, and complex LD structures across the genome, defining how signals from individual SNPs represent genes and further, components of pathways, has proved challenging.

Generally, two analysis strategies can be taken. One is to investigate a single, or a few, predefined candidate pathways for enrichment of significant p-values compared with all genes (Wang *et al.* 2009b). This can however introduce bias, both in gene selection and in assessing the significance of the finding compared with the untested pathways. The other is to use a hypothesis free approach and test all suggestively associated or differentially expressed genes, similar to the approach taken for example by Wang *et al.* (2007). This approach is biased by the data itself, i.e. it is dependent on the quality and properties of the initial GWAS or differential expression analyses, and should be interpreted within this context. However, focusing the analysis only on the most significant association hits is likely to ignore a significant number of false negative hits, especially in phenotypes such as autism, where large effect size variants are not present.

Currently, pathway analysis is greatly restricted by our limited knowledge of cellular processes and the far-from-complete functional annotation of genes. Yet, especially in the case of GWAS datasets, pathway analysis can in part help to overcome the problem of replication with small effect size associations, in particular in datasets with limited statistical power, in which single-locus genome-wide significance cannot be reached. Instead of focusing on individual genes with strongest evidence

of differential expression or association with the phenotype, pathway-based approaches typically rank all genes based on their significance and search for enrichment among the top end of the gene list (including the borderline significant results often ignored by GWA studies) thereby assessing whether seemingly scattered findings converge on the level of pathways.

Most of the pathway analysis methods for gene expression data are modifications of the original Gene Set Enrichment Analysis (GSEA) method, which determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states such as phenotypes (Mootha *et al.* 2003, Subramanian *et al.* 2005). The GSEA was used as the basis for one of the first pathway analysis algorithms developed for GWAS as well (Wang *et al.* 2007), which has been applied in many later studies (for e.g. Menashe *et al.*, Wang *et al.* 2009a).

Sequencing-based methods

Direct sequencing of DNA, i.e. traditional Sanger sequencing (Sanger and Coulson 1975), has typically been applied in candidate gene-based studies to screen for genetic changes in individual genes in limited numbers of samples. Due to the laborious and relatively expensive nature of sequencing, it is usually applied only to the coding, or exonic, parts of a gene. The increasing demand for low-cost and efficient sequencing lead to the recent development of high-throughput sequencing methods, usually referred to as next-generation sequencing. This technology, based on the parallel production of millions of sequence reads at once, has made it possible to sequence substantially larger regions of the genome faster and cheaper, although the storage, handling, and analysis of the data still imposes a substantial computational challenge.

With costs rapidly decreasing, it will soon be possible to sequence entire genomes of individuals for the purpose of research, as first initiated by the 1000Genomes project (www.1000genomes.org). The project aims to sequence the full genomes of ~1000 individuals from different global populations and provide a comprehensive resource on human genetic variation. As sequencing is a superior method to identify rare genetic variants, which are currently gaining increased attention within the disease genetics field, an intermediate approach of sequencing the coding regions of the genome ("exomes") has been taken whilst waiting for costs to plummet. Examples of studies where whole-exome sequencing has been successfully used to identify recessive mutations have already started to emerge (Bilguvar *et al.* 2010, Ng *et al.* 2010). In addition to DNA sequencing, next-generation sequencing can also been used for other applications. For example, RNA sequencing ("RNA-Seq") is used as a more comprehensive, probe-independent alternative to microarray-based gene

expression profiling, whereas "ChIP-Seq" is used to sequence immunoprecipitated DNA fragments in various epigenetic approaches.

2.1.6 Isolated populations

Isolated populations, also known as founder populations, are populations which have originated from a small number of original founders and subsequently rapidly expanded through normal population growth instead of immigration. Often these populations have experienced long periods of geographical or cultural isolation, which has further enhanced the isolation and lack of immigration. In such circumstances, genetic drift can have drastic effects on the gene pool by driving some alleles to fixation and others to extinction, thus reducing overall genetic heterogeneity. Since a low degree of genetic heterogeneity is ideal for all genetic mapping studies, the advantages of using isolated populations in disease genetics are well-recognized.

Finland is a well-known isolated population which has originated from a small number of original settlers. Compared with the rest of Europe, samples from the Finnish population tend to exhibit a decrease in genetic diversity (Sajantila *et al.* 1996) and increased degree of linkage disequilibrium (Varilo *et al.* 2003, Service *et al.* 2006) (see also Section 2.1.4), which has proved useful in mapping genetic diseases. Heterogeneity of the Finns is decreased also by various non-genetic factors, such as fairly uniform culture, lifestyle (such as diet), and standardized healthcare. Additionally, good genealogical records and reliable medical information make it easier to identify shared ancestry, construct pedigrees, and obtain correct diagnoses for the purpose of genetic studies. However, the distinct population history of Finland, characterized by multiple genetic bottlenecks followed by rapid growth without immigration, has left the population internally highly stratified, but also extremely homogenous within certain sub-isolates (Jakkula *et al.* 2008). Other isolated populations commonly used in genetic mapping studies include for example Iceland, Sardinia, Azores, and Newfoundland.

Due to its population history and strong founder effect, the Finnish population has a unique collection of enriched recessive disease alleles. These ~30 monogenic diseases are referred to as the Finnish disease heritage, and to date, at least one causative mutation has been successfully identified in all but one of them (Peltonen *et al.* 1999). On the contrary, some monogenic diseases which are common in Europe, for example cystic fibrosis and phenylketonuria, are very rare in Finland. In founder populations, these recessive diseases are often characterized by the presence of a single founder mutation, whereas numerous mutations in the same genes are identified in the global population. For example, in an autosomal recessive congenital nephrotic syndrome (CNS), 94 % of the Finnish cases carry one of two

mutations of the *NPHS1* gene, named FinMajor and FinMinor (Kestilä *et al.* 1998). Outside Finland, more than 119 different mutations in this gene have been described in CNS (Schoeb *et al.* 2010).

The processes behind the enrichment of recessive alleles have been postulated to affect risk alleles in common complex diseases as well, and the use of population isolates in complex disease genetic has quickly become popular. The tendency of the affected individuals to share ancestral haplotypes derived from a small number of founders can be expected to reduce overall noise in the analysis and increase detection power in genetic mapping studies (Peltonen *et al.* 2000). For example, large Finnish pedigrees were used successfully in a study of familial combined hyperlipidemia that identified the *USF1* gene (*upstream transcription factor 1)* as a risk factor for this complex disease (Pajukanta *et al.* 2004). Also, a gene conferring susceptibility to asthma (*NPSR1, neuropeptide S receptor 1*) was initially discovered by analyzing two regional subpopulations of Finland (Laitinen *et al.* 2004). More recently, common alleles affecting susceptibility for multiple sclerosis were identified in a high-risk Finnish sub-isolate (Kallio *et al.* 2009, Jakkula *et al.* 2010). An important observation from these studies is that even if genetic variants predisposing to complex traits are originally identified from an isolated population or exceptional large families, they are often replicated in large-scale population samples elsewhere (Kristiansson *et al.* 2008).

The era of genome-wide association studies has greatly increased the general awareness of population stratification, i.e. genetic differences among populations. With up to few million analyzed markers and study samples reaching tens of thousands, even slight effects of stratification can introduce huge bias to the results. The International HapMap Project (2005) was among the first to assess allele frequency differences of common variants among the Caucasian, African, Chinese, and Japanese populations , and subsequently, the genetic variation among many global populations has been extensively characterized (for e.g. Abdulla *et al.* 2009, Tishkoff *et al.* 2009, Xu *et al.* 2009, Behar *et al.* 2010). It is now known that significant population substructure can be present also within populations, such as in Europe (Novembre *et al.* 2008, Salmela *et al.* 2008). Strikingly, even in isolated populations such as Finland, substantial genetic substructure has been detected (Jakkula *et al.* 2008).

2.1.7 Micro-RNAs

Although GWA studies and other genetic mapping approaches have robustly identified a large number of loci and specific variants which contribute to the risk of many common diseases, the relationship between these variants and the actual disease mechanisms remains mostly unknown, despite a few exceptions (for e.g.

Moffatt *et al.* 2007). The detection and genome-wide mapping of eQTLs in different human tissues (Stranger *et al.* 2007a, Dimas *et al.* 2009) has greatly promoted the conception of what such mechanisms could be like, and further emphasized the importance of characterizing different types of regulatory elements in the genome. Transcription factors (TF) have traditionally been the key players in regulatory network studies, but the discovery of micro-RNAs (miRNA) in 2001 added another layer of complexity to these studies. Since both TF and miRNA binding sites can be disrupted by SNPs or CNVs, they have quickly become the target for intensive research in human genetics. For the purpose of Study I of this thesis, some of the central concepts of miRNAs will be covered in this review.

Biogenesis

Micro-RNAs (miRNAs) are short, ~22 nucleotide-long non-coding RNA-molecules, which occur naturally in cells and are capable of silencing protein coding genes post-transcriptionally by specifically binding the messenger RNA (mRNA). Micro-RNAs are processed from longer transcripts which are initially either transcribed from independent miRNA genes by RNA polymerase II (pri-miRNAs), or processed from introns of protein-coding genes ("mirtrons") (Kim *et al.* 2009). Thus, a mature miRNA can occur from multiple distinct genomic loci, i.e. multiple different primary transcripts. An overview of the miRNA biogenesis is presented in Figure 2. The pri-miRNAs fold into hairpin structures, which are cleaved by an RNase-enzyme, Drosha, into ~70-nucleotide pre-miRNAs. These pre-miRNAs are then exported from the nucleus into the cytoplasm where they are further cleaved by another RNase, Dicer, into ~20bp RNA duplex. One strand of this duplex represents the mature miRNA, which is subsequently loaded into a miRNA-induced silencing complex (miRISC). The other strand is called the "passenger" strand and is often, but not always, degraded. If both strands remain, the less abundant form (the "passenger") is denoted with an asterisk (*) in miRNA nomenclature. If predominance cannot be determined, the two forms are denoted with "3p" and "5p", referring to the 5’ and 3’ arms of the stem-loop precursor. The actual silencing process is carried out by the miRISC complex, together with associated molecules such as argonaute proteins, when the mature miRNA base-pairs with the target mRNA (Krol *et al.* 2010).

Silencing mechanism

The miRNA-mediated silencing effect is achieved by either suppressing the translation or initiating the degradation of the target mRNA. Whilst the mechanism of mRNA deadenylation followed by its decay is fairly well characterized, the process of translational repression remains poorly understood (for e.g. Fabian *et al.* 2010). The relative contributions of these two outcomes have been largely unknown,

particularly for endogenous targets which are expressed at relatively low levels and are hard to measure accurately. However, a recent study demonstrated that lowered mRNA levels account for 84% of the decreased protein production, and changes in mRNA levels closely reflect the impact of miRNAs on gene expression (Guo *et al.* 2010). The authors also suggest that destabilization of target mRNAs is the predominant reason for reduced protein output. Mature miRNAs recognize their targets through specific basepairing between 5' nucleotides 2 – 8 of the miRNA (which is referred to as the "seed" region) and complementary nucleotides in the 3' untranslated region (3'UTR) of the target mRNA (Lewis *et al.* 2005). It is known that transcripts with multiple, non-overlapping miRNA binding sites are more responsive to miRNA-induced repression than those with a single binding site (Doench *et al.* 2003, Zeng *et al.* 2003). Interestingly, recent findings have also shown that under certain conditions, or in specific cells for example, miRNA-mediated repression can be reversed or prevented. Further, interaction of the silencing complex and the mRNA 3'UTR can lead to upregulation rather than downregulation of translation (Vasudevan and Steitz 2007), which adds a whole new dimension to the regulatory capabilities of miRNAs.

Micro-RNA target prediction and relevance to disease

Micro-RNA genes are abundant in the genome and it has been estimated that they regulate at least 50% of all human genes (Krol *et al.* 2010). The central miRNA database, miRBase (http://www.mirbase.org) (Griffiths-Jones *et al.* 2006), currently contains 940 annotated human miRNAs (release 15). The recognition that the specificity of the miRNA - target mRNA recognition arises from base-pairing between the seed region and the 3'UTR had direct implications for disease genetics. Since the target sites are subject to disruption by SNPs or mutations as any other genomic sequence leading to abnormal miRNA regulation of the gene or transcript in question, they can mediate disease susceptibility through variation among individuals. A multitude of miRNA target prediction algorithms subsequently followed and have been widely applied since. However, accurate prediction has proved challenging due to the extremely short length of the seed sequence, and the number of false-positive predictions is huge, which is why functional validation of targets is constantly increasing.

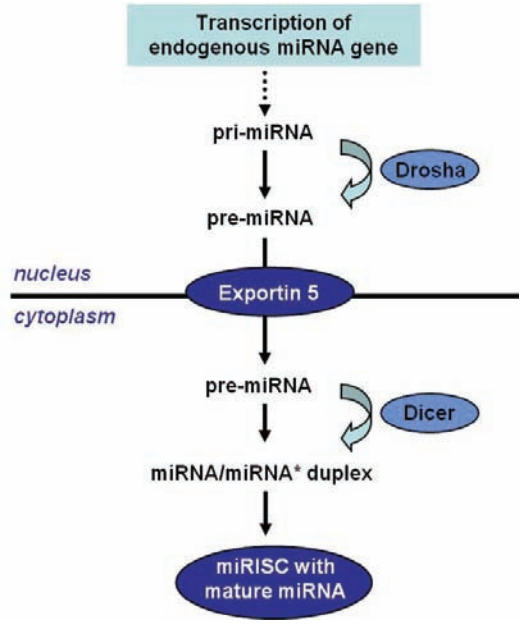


Figure 2. ***Outline of the micro-RNA biogenesis.***

Most of the prediction algorithms are based on a few major criteria. First, the predictions are currently limited to the 3'UTR sequences, although it is possible that target sites are present also elsewhere. Second, the exact base-pairing between the mRNA and the 5' region of the miRNA is most crucial with seed nucleotides 2 – 7, but prediction specificity increases, when an 8-nucleotide match is required. It is also known that the length of the complementary region correlates with the efficiency of silencing (Grimson *et al.* 2007, Bartel 2009). Third, evolutionarily conserved sites are more likely to be true sites, and highly conserved miRNAs have many conserved targets (Lewis *et al.* 2005, Xie *et al.* 2005), even though 3'UTR sequences are typically poorly conserved. Some of the most commonly used prediction algorithms include TargetScan (www.targetscan.org), miRanda (www.microrna.org), and PicTar (www.pictar.org). Additional tools, such as RNAhybrid (http://bibiserv.techfak.uni-bielefeld.de/rnahybrid) and RNA22 (http://cbcsrv.watson.ibm.com/rna22_targets.html), can take into account the thermodynamical properties of the binding between the miRNA and the target mRNA n their predictions. Common practice is to use multiple algorithms to do a prediction and then rank the obtained results based on convergent evidence from different methods. Target prediction methodology is additionally covered in Section 4.3.2.

There is an increasing amount of evidence of the involvement of miRNAs in disease. The most common approach is a genome-wide analysis of differential miRNA expression in cases and controls, but targeted functional studies of individual miRNAs are also emerging (for e.g. Sethupathy *et al.* 2007, Tan *et al.* 2007, Hollander *et al.* 2010). Although most studies have been performed in cancer (for e.g. Lu *et al.* 2005), there is also evidence of miRNA involvement in neuropsychiatric and neurodegenerative phenotypes, such as schizophrenia (Kim *et al.* 2010) and Alzheimer's disease (Wang *et al.* 2008).

One of the first and well-known examples of miRNA involvement in disease was from Tourette syndrome (TS), a condition phenotypically related to ASDs (Abelson *et al.* 2005). In this study, a frameshift mutation and two independent occurrences of the same non-coding sequence variant were identified in the 3'UTR of the *SLITRK1* gene in 174 unrelated TS probands. The variants were located in a conserved binding site for hsa-miR-189, and the authors showed that in the presence of hsa-miR-189, the variants reduced the expression of *SLITRK1* compared with the wild-type gene. The variant was absent from 4296 controls, leading to a statistically significant association with TS. The authors further demonstrated that the expression patterns of *SLITRK1* and hsa-miR-189 were overlapping and developmentally regulated in both mouse and human brain regions previously implicated in TS, and that the normal function of *SLITRK1* in dendritic growth in primary neuronal cultures was affected by the mutation. Although this example only applies to rare cases of TS, it created a lot of excitement, because it was able to demonstrate a mechanism in humans, through which miRNAs can affect a disease phenotype. However, as with all mechanisms regulating gene expression, tissue-specificity of miRNAs remains an important question. Especially in disease-oriented miRNA studies like this one, addressing tissue specificity will be challenging since most miRNAs are likely have spatially and temporally narrow regulatory effects.

2.2 Autism Spectrum Disorders

2.2.1 Clinical features

Autism spectrum disorders (MIM#209850), also known as pervasive developmental disorders (PDDs) (F84), include childhood autism (or autistic disorder, AD) (F84.0), Asperger syndrome (AS) (F84.5), atypical autism (F84.1), childhood disintegrative disorder (F84.3), and Rett syndrome (F84.2) according to the current International Classification of Diseases (ICD-10, World Health Organization 1993). An additional, commonly used term is PDD-NOS (pervasive developmental disorder not otherwise specified) (F84.9), which refers to an "autism-like" phenotype similar

to childhood autism, but falling short of the strict diagnostic criteria. ASDs have traditionally been conceptualized by most researchers as a continuum of the same disorder with varying degrees of severity, and in fact, it has been suggested that in the upcoming new edition of the ICD-classification, the disorders would be treated as a continuum of phenotypes. It is also likely that Rett syndrome, a close-to-monogenic form of ASDs affecting predominantly girls (Amir *et al.* 1999), will be excluded from the group.

Childhood autism is characterized by qualitative impairment or delayed development in verbal and non-verbal communication, reciprocal social interaction, and behavioural skills before the age of three years. The extent and quality of symptoms can vary significantly, and the impairments can be expressed in different ways. For example, although delay or absence of spoken language is fundamental in autism, it is not present in all affected individuals (Alarcon *et al.* 2002). Similarly, some affected individuals clearly avoid all forms of social interactions, whereas others seek actively for personal interactions, even if in that would be in a socially odd manner (Wing and Gould 1979, Volkmar *et al.* 1989). Asperger syndrome is a milder disorder of the spectrum, which shares the core clinical features of childhood autism, but does not present any major cognitive deficiencies. Individuals with AS typically have a fairly normal language development and average basic verbal skills for communication, and the disorder is usually recognized much later than childhood autism (Volkmar and Klin 2000). Typical features of AS include difficulties in socialization, one-sided way of communication, unusual patterns of interest, formal and pedantic speech and dependence of routines and rituals (Nieminen-Von Wendt 2004). All diagnoses of ASDs are based on structured interviews and behavioral observation, in the absence of any molecular or physiological markers. The most commonly used diagnostic tool is the Autism Diagnostic Interview –Revised (ADI-R) (Lord *et al.* 1994) accompanied by the Autism Diagnostic Observational Schedule (ADOS) (Lord *et al.* 1989), which have quickly become the gold standard tools in most autism research. The diagnostic criteria of childhood autism and AS, the two central phenotypes in this study are presented in Tables 1 and 2.

Table 1. ***ICD-10 diagnostic criteria for childhood autism (F84.0) (World Health Organization).***

A. Presence of abnormal or impaired development before the age of three years, in at least one out of the following areas:
(1) receptive or expressive language as used in social communication;
(2) the development of selective social attachments or of reciprocal social interaction;
(3) functional or symbolic play.

B. Qualitative abnormalities in reciprocal social interaction, manifest in at least one of the following areas:
(1) failure adequately to use eye-to-eye gaze, facial expression, body posture and gesture to regulate social interaction;
(2) failure to develop (in a manner appropriate to mental age, and despite ample opportunities) peer relationships that involve a mutual sharing of interests, activities and emotions;
(3) A lack of socio-emotional reciprocity as shown by an impaired or deviant response to other people's emotions; or lack of modulation of behaviour according to social context, or a weak integration of social, emotional and communicative behaviours.

C. Qualitative abnormalities in communication, manifest in at least two of the following areas:
(1) a delay in, or total lack of development of spoken language that is not accompanied by an attempt to compensate through the use of gesture or mime as alternative modes of communication (often preceded by a lack of communicative babbling);
(2) relative failure to initiate or sustain conversational interchange (at whatever level of language skills are present) in which there is reciprocal to and from responsiveness to the communications of the other person; (3) stereotyped and repetitive use of language or idiosyncratic use of words or phrases;
(4) abnormalities in pitch, stress, rate, rhythm and intonation of speech;

D. Restricted, repetitive, and stereotyped patterns of behaviour, interests and activities, manifest in at least two of the following areas:
(1) an encompassing preoccupation with one or more stereotyped and restricted patterns of interest that are abnormal in content or focus; or one or more interests that are abnormal in their intensity and circumscribed nature although not abnormal in their content or focus.
(2) apparently compulsive adherence to specific, non-functional, routines or rituals;
(3) stereotyped and repetitive motor mannerisms that involve either hand or finger flapping or twisting, or complex whole body movements;
(4) preoccupations with part-objects or non-functional elements of play materials (such as their odour, the feel of their surface, or the noise or vibration that they generate);
(5) distress over changes in small, non-functional, details of the environment.

E. The clinical picture is not attributable to the other varieties of pervasive developmental disorder; specific developmental disorder of receptive language (F80.2) with secondary socio-emotional problems; reactive attachment disorder (F94.1) or disinhibited attachment disorder (F94.2); mental retardation (F70-F72) with some associated emotional or behavioural disorder; schizophrenia (F20) of unusually early onset; and Rett's syndrome (F84.2).

Table 2. ***ICD-10 diagnostic criteria for Asperger Syndrome (F84.5) (World Health Organization).***

A. A lack of any clinically significant general delay in spoken or receptive language or cognitive development. Diagnosis requires that single words should have developed by two years of age or earlier and that communicative phrases be used by three years of age or earlier. Self-help skills, adaptive behaviour and curiosity about the environment during the first three years should be at a level consistent with normal intellectual development. However, motor milestones may be somewhat delayed and motor clumsiness is usual (although not a necessary diagnostic feature). Isolated special skills, often related to abnormal preoccupations, are common, but are not required for diagnosis.

B. Qualitative abnormalities in reciprocal social interaction (criteria as for autism).

C. An unusually intense circumscribed interest or restricted, repetitive, and stereotyped patterns of behaviour, interests and activities (criteria as for autism; however it would be less usual for these to include either motor mannerisms or preoccupations with part- objects or non-functional elements of play materials).

D. The disorder is not attributable to the other varieties of pervasive developmental disorder; schizotypal disorder (F21); simple schizophrenia (F20.6); reactive and disinhibited attachment disorder of childhood (F94.1 and .2); obsessional personality disorder (F60.5); obsessive-compulsive disorder (F42).

Intellectual disability (ID), also referred to as mental retardation, and epilepsy are the most commonly occurring associated medical conditions in ASDs. ID occurs in 75-80% and seizures in 25-30% of autistic children (Bailey *et al.* 1996, Fombonne 1999, Gillberg and Billstedt 2000), although a more recent study estimated the prevalence of cognitive defects to be lower than 50% (Chakrabarti and Fombonne 2005). Information regarding AS is less well documented. About 10–15% of individuals with childhood autism have co-occurring medical conditions with known etiologies (Folstein and Rosen-Sheidley 2001), which are usually specific cytogenetic or single gene disorders. The most common associated Mendelian disorders are Fragile X syndrome (MIM#300624) (in 1-2% of ASD cases), tuberous sclerosis (TSC; MIM#191100) (in ~1% of ASD cases), and chromosome 15q11-13 duplication syndrome (in 1-2% of ASD cases), which are reviewed for example by Gillberg and Billstedt (2000) or Freitag (2007) along with other coexisting disorders. The wide range of associated medical conditions in ASDs further adds to its clinical heterogeneity, and highlights the importance of accurate phenotyping for genetic studies, since the underlying genetic background is likely to reflect this heterogeneity, even in the idiopathic forms of the disorder. Also, as suggested, it is probably more appropriate to think about "the autisms" instead of arbitrarily grouping all existing cases under a single diagnostic class (Geschwind and Levitt 2007) (Figure 3).

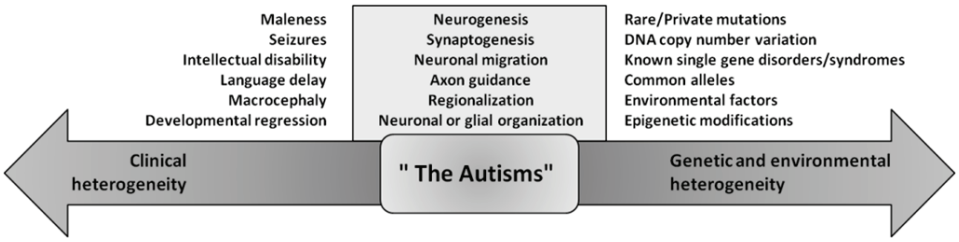


Figure 3. ***Illustration of the concept of "The Autisms", emphasizing some of the many etiological features and the clinical heterogeneity within ASDs.*** *The central box lists some of the neurodevelopmental processes that might be disrupted in ASDs. Figure adapted from Geschwind and Levitt (2007).*

2.2.2 Prevalence

Based on the most recent epidemiological review on ASDs (Fombonne 2009), the overall prevalence of all ASDs is 60-70/10 000 (0.6-0.7% of children), making it one of the most frequent childhood neurodevelopmental disorders. This study reviewed the results of altogether 43 studies published since 1966, and provides very good estimates of the current figures. The prevalence of strictly defined childhood autism is ~20/10 000, whereas for AS, the review gives an estimate of ~6/10 000 which should be treated with caution, since only a few small epidemiological surveys specific for AS have been conducted (Ehlers and Gillberg 1993, Kadesjo *et al.* 1999, Mattila *et al.* 2007). Thus, this estimate is mostly based on more recent autism surveys, in which AS has been additionally studied. Though AS was first described already in 1944 (Asperger 1944), it remained widely unrecognized until the 1980's, and was added to the ICD-10 as an independent diagnosis as late as 1993, which explains the small number of surveys thus far. However, overall, the prevalence of AS seems much lower than that of childhood autism. The prevalence of PDD-NOS was estimated to be around 30/10 000 (Fombonne 2009).

All ASDs are generally more common in males than in females (ratio ~4:1), but this difference is much smaller when only severe forms of childhood autism with intellectual disability are considered. On the contrary, in high-functioning children with childhood autism, the male-to-female ratio can be as high as 8:1 (Fombonne 2005). The sex bias has not been explained to date, but it seems that it is not driven by X-chromosomal loci (Abrahams and Geschwind 2008) (see also Section 2.2.8). An exception to this is Rett syndrome, which occurs almost exclusively in females and is caused by mutations in *Methyl-CpG-binding protein 2* (*MeCP2*) at chromosome Xq28 in ~80% of cases (Amir *et al.* 1999).

The prevalence of all ASDs has notably increased in the last decades, and older studies systematically obtain lower prevalence estimates than the more recent ones. Various environmental reasons for this have been proposed (reviewed for e.g. in Landrigan 2010), such as maternal exposure to rubella virus, but the majority of the field attributes the increase mainly to increased awareness of the disorders and broader diagnostic criteria rather than actual increased incidence. In particular, awareness of AS and milder phenotypes of the spectrum, such as PDD-NOS, has greatly increased and diagnostic accuracy has improved. Also, advances in health services worldwide have probably brought more affected individuals into surveys.

2.2.3 Mode of inheritance

In the 1940's, both Kanner and Asperger noticed ASD-like personality traits in the parents of affected children in their original studies, and made a preliminary suggestion of a heritable component in the disorders they described. Still, it took a long time before the genetic and biological basis of autism was widely recognized and accepted. In the 1960's it was thought that autism was caused by poor parenting, in particular mothers withholding their affection. Even though Kanner himself said in the early years that parental coldness might contribute to autism (Kanner 1949), he later renounced this theory. Yet, his statement was misused for a long time, and the concept of "refridgerator mothers" prevailed until the biological basis of autism was eventually established (Rutter 1968). The importance of genetic factors became clear when the co-occurrence of autism with chromosomal aberration syndromes, such as Fragile-X, was noticed (Blomquist *et al.* 1985).

According to twin studies, childhood autism is clearly heritable. Concordance rates vary from 69-98% among MZ twins to 0-30% in DZ twins (Folstein and Rutter 1977, Steffenburg *et al.* 1989, Bailey *et al.* 1995). However, most of the twin studies have been small, looking at tens of twin pairs only. In family studies, 2-6% of siblings of individuals with autism were found to have an ASD (Bailey *et al.* 1998). The heritability estimate, calculated from the MZ:DZ concordance ratio and sibling recurrence risk, is ~90%, which is one of the highest among complex disorders, (Szatmari *et al.* 1998, Folstein and Rosen-Sheidley 2001).

For AS, no systematic twin studies have been performed, and the estimation of familial aggregation has been challenging due to the availability of case reports only. However, as noticed already by Asperger himself, family members of individuals with AS often have problems with social interaction as well, suggesting the involvement of genetic factors (Asperger 1944, Bowman 1988, Gillberg 1989, Volkmar *et al.* 1998). Also, it should be noted that even though AS is often seen in siblings of individuals with autism, large families exist where AS is transmitted

through the pedigree, in the absence of childhood autism, in a manner resembling dominant mode of inheritance (Ylisaukko-oja *et al.* 2004, Rehnström *et al.* 2006).

The genetic basis of autism has traditionally been thought to resemble that of other common, complex disorders, with multiple common susceptibility variants comprising the majority of the risk. Earlier, it was estimated that the number of interacting loci contributing to autism susceptibility ranged between two and 15 genes of varying effect (Pickles *et al.* 1995, Risch *et al.* 1999), but in the light of the knowledge gained from for example genome-wide association studies, the true number of involved loci is likely to be in hundreds. Also, recent GWA studies have shown that effect sizes for common variants in ASDs are exceptionally low (see Section 2.2.6), suggesting that rare genetic events are more likely to play a significant role in ASDs. Concurrent reports of rare, *de novo* CNVs, and rare, high penetrant mutations identified in individual ASD families (Section 2.2.7 and 2.2.9) have further strengthened this view, and underscored the substantial genetic heterogeneity in ASDs. Given that not all variants are fully penetrant and their expressivity can vary, the interaction between rare and common genetic variants needs to be fully characterized in order to understand the underlying genetic model(s) in ASDs. Also, it should be remembered that the autistic phenotype is not only a product of interactions between different combinations of susceptibility variants, but also of CNVs and other chromosomal aberrations, epistatic mechanisms, and epigenetic and environmental factors.

2.2.4 Biological basis

As with genetic findings, the underlying neurobiology of ASDs is likely to be heterogeneous and no single hypothesis has gained more support than others. ASDs are considered "developmental" disorders because symptoms appear soon after birth in early infancy, and they affect many aspects of cognition and behaviour, leaving them poorly developed. Still, it is completely unknown at which point the primary lesion (whatever that is) occurs, and why the affected processes specifically relate to language and social functions.

The neuropathological findings in autism are greatly complicated by frequent co-morbid features (see Section 2.2.1). Numerous small abnormalities have been reported, such as enlargement of the hippocampus and the amygdala (Schumann *et al.* 2004), but these have mostly been inconsistent. Postmortem and structural magnetic resonance imaging studies have highlighted the frontal lobes, amygdala and cerebellum as pathological (Amaral *et al.* 2008), and a number of functional neuroimaging studies have suggested that impairments in cortical connectivity are present in individuals with autism (Just *et al.* 2004, Koshino *et al.* 2008). In one of these studies, the brain activation of a group of high-functioning autistic participants

and intelligence quotient (IQ) matched controls was measured using functional magnetic resonance imaging (fMRI) during sentence comprehension. The authors report that the functional connectivity, meaning the degree of synchronization of the time series of the activation between the various participating cortical areas, was consistently lower for the autistic than the control participants (Just *et al.* 2004).

One of the main biological themes in ASDs that has emerged is synaptic dysfunction. This was initially suggested by rare, high-penetrance mutations associated with autism which have recently been identified in multiple genes encoding for synaptic cell adhesion molecules, such as neuroligins, neurexins, and *contactin-associated protein-like 2* (*CNTNAP2*) (see Section 2.2.9). These molecules connect pre- and post-synaptic neurons at the synaptic cleft, mediate synaptic signalling, and are suggested to shape neural networks by specifying synaptic functions (Sudhof 2008). Another suggested theme is neuronal excitability, i.e. the ratio of excitation and inhibition, which seems to be influenced by a number of autism-associated genes (Rubenstein and Merzenich 2003) such as neuroligins, whose mutations appear to affect the balance between excitatory and inhibitory synaptic transmission (Tabuchi *et al.* 2007). As suggested by Walsh and colleagues (2008), a unifying hypothesis might be that genes involved in autism would encode proteins which mediate activity-dependent changes of neuronal function (Hong *et al.* 2005). Thus, autism would reflect abnormal regulation of gene expression under specific, neuronal activity-dependent processes, such as learning. This hypothesis is supported by a recent study, where homozygosity mapping in consanguineous autism families revealed homozygous deletions affecting genes whose level of expression changes in response to neuronal activity (Morrow *et al.* 2008). However, it is unclear how dysfunction in such general synaptic function could lead to autism and yet leave so many cognitive processes unaffected. Interestingly, it is known that early stages of neurodevelopment, such as neuronal production, axonal growth, and initial connectivity, are largely independent of synaptic functioning (Sur and Rubenstein 2005). Instead, the refinement of synapses, which occurs at a later developmental stage, relies on neuronal activity and eventually leads to synaptic plasticity and learning (Hong *et al.* 2005).

Other biological processes proposed to be affected in ASDs include glutamatergic and serotonergic neurotransmission (Bear *et al.* 2004, Chugani 2004), and abnormal calcium signalling (Krey and Dolmetsch 2007). However, the genetic evidence for synaptic dysfunction and impaired synaptic cell adhesion coupled with the structural and functional evidence of cortical underconnectivity has lead to the prevailing hypothesis that autism is a neuronal disconnection syndrome (Frith 2004, Courchesne and Pierce 2005, Geschwind and Levitt 2007, Hughes 2007, Wang *et al.* 2009b).

2.2.5 Linkage studies

After the heritability and the genetic basis of ASDs was established, the next step was to locate the regions of the genome where the predisposing genes and genetic factors would reside. Microsatellite-based genome-wide linkage analysis together with regional candidate gene studies was the method-of-choice in autism genetics, as well as in other complex disease phenotypes, for many years until genome-wide SNP approaches emerged. The objective of a linkage analysis is to find a locus in the genome that is inherited together with the trait in question more often than it should by chance. Numerous studies have been performed in different populations, but overall, most identified loci have reached only suggestive levels of significance and replication has been marginal, as in many other complex diseases (Altmuller *et al.* 2001). Study samples have been small, especially in the early days, but it has also been shown that increasing the study sample does not necessarily result in comparably more significant linkage signals (Yonan *et al.* 2003). The lack of genome-wide significant linkage and replication probably reflects the underlying heterogeneity of the used study samples. Especially in large collaborative studies, samples are often from all over the world, increasing both genetic and diagnostic heterogeneity. To decrease heterogeneity, efforts have been made to subgroup study samples based on various selected phenotypic features, such as sex or age at first word. This approach has improved signals and even revealed a few loci reaching genome-wide significance, once again highlighting the challenges heterogeneity poses on all genetic research. Most of the linkage studies performed in ASDs have been extensively reviewed recently (Freitag 2007, Yang and Gill 2007, Abrahams and Geschwind 2008), and a summary of the most consistent findings is presented in Table 3. In this review, I will only focus on some of the most important findings, and findings from the Finnish population.

Linkage signals for ASDs have been reported in almost all chromosomes, and no single locus can really be picked out as more significant than others, which is typical for complex neuropsychiatric disorders. Chromosome seven is probably the most studied chromosome in ASDs, and the only locus which was supported by two meta-analyses (Badner and Gershon 2002b, Trikalinos *et al.* 2006). Linkage has been observed at two distinct loci, 7q22-q31 and 7q34-q36 (IMGSAC 1998, Ashley-Koch *et al.* 1999, Barrett *et al.* 1999, IMGSAC 2001a, b, Liu *et al.* 2001, Shao *et al.* 2002, Lamb *et al.* 2005), which both harbour some of the most rigorously studied candidate genes in ASDs, such as *CNTNAP2*, *RELN*, and *MET* (see Section 2.2.9). Chromosome 7q34-q36 has also been implicated by QTL-based linkage studies using endophenotypes instead of the end-state diagnosis. In studies by Alarcon and colleagues (Alarcon *et al.* 2002, Alarcon *et al.* 2005), three quantitative measures from the ADI-R screening questionnaire were considered, and evidence of linkage was found for language delay (age at first word) at this locus.

However, in the two largest linkage studies in ASDs to date (Yonan *et al.* 2003, Szatmari *et al.* 2007), support for linkage at chromosome seven was not observed. In the study by Yonan and colleagues, 345 multiplex families were analyzed, each with at least two siblings affected with ASDs. The most significant findings were on 17q11 (maximum LOD score [MLS] = 2.83, p =0.00029) and on 5p13 (MLS=2.54, p =0.00059). The chromosome 17q locus is near the serotonin transporter (5-HTT) gene *SLC6A4*, and linkage evidence for this locus has been observed in at least five other studies (IMGSAC 2001b, Bartlett *et al.* 2005, Cantor *et al.* 2005, McCauley *et al.* 2005, Trikalinos *et al.* 2006). What makes this locus especially interesting is that three studies found considerable evidence of sex-specific effects at this locus. All of these studies divided their study samples to male-only versus female-only containing sibships, and in all cases, significant linkage was observed at 17q11 with the male-only families (Stone *et al.* 2004, Cantor *et al.* 2005, Sutcliffe *et al.* 2005). However, in a follow-up study, no single SNP or haplotype was sufficient to account for the linkage signal (Stone *et al.* 2007).

In the study by Szatmari and colleagues (2007), which reported the results of the Autism Genome Project Consortium (AGP), 11p12-p13 was the single major locus identified. Altogether 1181 families were analyzed for linkage and DNA copy number variation. Based on the CNV analyses, further subsetting of the families was performed to decrease heterogeneity, and suggestive linkage evidence was observed also for 15q23–25.3, in addition to 11p12–p13. This still did not change the fact that even with the largest reported ASD study sample thus far, almost no overlap was observed with previously published studies.

The first genome-wide linkage scan for ASDs in the Finnish population was published in 2002 (Auranen *et al.* 2002). The authors analyzed 38 multiplex families with 87 affected individuals. The most significant LOD scores were observed at chromosome 3q25-27 (Z_{max}=4.31, MLS 4.81). Two other loci at 1q21-23 and 7q also showed some evidence of linkage (lod scores > 2). A follow-up study was conducted where these three loci were finemapped, but overall, it did not add much to the significance of the original study (Auranen *et al.* 2003).

Another independent genome-wide linkage scan in Finland was performed in Asperger syndrome (Ylisaukko-oja *et al.* 2004). This study is still the only linkage study specifically for AS. Seventeen large pedigrees with 82 AS cases (and no autism cases) in multiple subsequent generations were analyzed. Strongest linkage was observed at chromosomes 1q21-23 (Z_{max}=3.58), 3p13-24 (Z_{max}=2.50), and 13q31-33 (Z_{max}=1.59), and the linkage to the 3p locus was later replicated in an independent set of 12 Finnish AS families (Rehnström *et al.* 2006). Linkage to 1q21-23 and 13q31-33 has been observed also in multiple linkage studies for schizophrenia (Blouin *et al.* 1998, Brzustowicz *et al.* 1999, Brzustowicz *et al.* 2000, Gurling *et al.* 2001, Badner and Gershon 2002a).

Table 3. ***Summary of ASD linkage peaks with support from multiple studies.*** *Listed are loci, which obtained a logarithm of odds (LOD) score > 3 in at least one study, and a LOD > 2 in at least one additional study. The studies indicated with a star (*) performed also an analysis with families with affected males only. "Age at first word" is a measure taken from the ADI-R diagnostic interview, whereas the SRS screening tool measures the severity of social impairment associated with ASDs, and is completed by parents and/or teachers. Table adapted from Abrahams and Geschwind (2008).*

Locus	Phenotype	Cohort	Reference
1q21-q23	AS	FIN	Ylisaukko-oja *et al.* (2004)
	ASD	FIN	Auranen *et al.* (2002)
2q23-q31	ASD	IMGSAC	IMGSAC (2001b)
	ASD	US	Buxbaum *et al.* (2001)
3q25-q27	ASD	FIN	Auranen *et al.* (2002)
	ASD	Utah	Coon *et al.* (2005)
5p13-p14	ASD	AGP	Szatmari *et al.* (2007)
	ASD	AGRE	Liu *et al.* (2001)
	ASD	AGRE	Yonan *et al.* (2003)
7q22-q31	ASD	IMGSAC	IMGSAC (2001b)
	ASD	IMGSAC	IMGSAC (1998)
7q34-q36	Age at first word	AGRE	Alarcon *et al.* (2002)
	Age at first word	AGRE	Alarcon *et al.* (2005)
9q33-q34	Age at first word	CPEA	Schellenberg *et al.* (2006)
	ASD	AGP	Szatmari *et al.* (2007)*
11p12-p13	ASD	AGP	Szatmari *et al.* (2007)
	SRS score	AGRE	Duvall *et al.* (2007)
17q11-q21	ASD	AGRE	Stone *et al.* (2004)*
	ASD	AGRE	Cantor *et al.* (2005)*
	ASD	AGRE	Yonan *et al.* (2003)

ABBREVIATIONS: AS=Asperger syndrome, ASD=autism spectrum disorder, FIN=Finland, IMGSAC=International Molecular Genetic Study of Autism Consortium, AGRE=Autism Genetic Resource Exchange, CPEA=Collaborative Programs of Excellence in Autism Network at the National Institute of Health, AGP=Autism Genome Project, SRS=Social Responsiveness Scale

In summary, no consistent linkage evidence for any chromosomal region has been found in ASDs. Even when loci are relatively well replicated in multiple smaller studies, they have not been detected by the larger ones at all, further emphasizing the heterogeneity of the disorder and the need for even larger study samples. Also, since many of the published linkage studies have used overlapping or even same family material, their evidence of linkage cannot be considered independent. However, in the light of what is currently known about the genetic background of autism, there is a good chance that many of the identified linkage loci are real and that they harbor either multiple rare variants, possibly specific for a subset of samples, or common

variants with very small effects which escape consistent detection due to insufficient sample size.

2.2.6 Genome-wide association studies

To date (August 2010), three genome-wide association studies (GWAS) in autism have been published. The study by Wang and colleagues (2009b), reported the first and largest-to-date autism GWAS, with multiple large ASD cohorts. They started by analyzing 780 families from the Autism Genetic Resource Exchange (AGRE) with 1299 affected individuals. No genome-wide significant signals were detected so they analyzed the Autism Case-Control (ACC) Cohort with 1204 additional cases (2503 cases in total), and performed a meta-analysis with the AGRE data. This analysis yielded one genome-wide significant hit at chromosome 5p14.1 (rs4307059, $p=3.4x10^{-8}$). To replicate this finding, two additional cohorts with 540 and 108 cases were analyzed. In both replication cohorts, the most significant association signals from the discovery cohorts were replicated with the same direction of association. The most significant SNPs all tagged the same haplotype block between two cadherin genes, *CDH9* and *CDH10*, both of which mediate calcium dependent cell-cell adhesion. A pathway-based association analysis further supported the role of neuronal cell adhesion molecules in autism susceptibility, which is in line with previous genetic findings. It should be noted that a separate GWAS paper has been published where one of the replication cohorts was analyzed independently, and the AGRE samples used as a replication cohort (Ma *et al.* 2009). With far less convincing evidence, the authors report common variants at the same 5p14.1 locus, which is not surprising, given that the study sample is completely overlapping with the Wang *et al.* study.

The second study, by Weiss and colleagues (2009), performed a family-based study of genome-wide linkage and association in 1031 multiplex autism families with altogether 1553 affected individuals. All families had at least one affected individual meeting the ADI-R criteria for autism, and possible family members affected with an ASD were included. Families were obtained from the AGRE and US National Institute for Mental Health (NIMH) repositories. The study identified two regions of either suggestive or significant linkage at 6q27 and 20p13, respectively. As with the Wang *et al.* study, no genome-wide significant associations were identified with the initial study sample, but after the authors proceeded to genotype the most significant SNPs in additional autism families, one SNP at 5p15 reached genome-wide significance ($p=2x10^{-7}$) in the combined analysis. The SNP was located between two genes, *SEMA5A* and *TAS2R1*, and due to previous evidence of *SEMA5A* in ASDs, the authors continued to demonstrate that the expression of *SEMA5A* was reduced in cortical samples of individuals with autism, thus confirming a previously reported finding from LCLs (Melin *et al.* 2006).

The third study reported results from the Autism Genome Project (AGP) Consortium GWAS with 1369 ASD families (1385 affected individuals) (Anney *et al.* 2010). One SNP, rs4141463 at chromosome 20p12.1, exceeded genome-wide significance ($p=2.1x10^{-8}$). The authors proceeded to analyze a smaller replication study sample (595 families with 1086 affected individuals), again from the AGRE, but this did not add to the significance of the results, and in the combined analysis of both datasets, rs4141463 barely reached genome-wide significance ($p=4.7x10^{-8}$). The SNP is located in *MACROD2* gene, which is largely unknown. Additional exploratory analyses with phenotype subgroups were also performed, but no significant association results were obtained from these, considering the burden of multiple testing.

Importantly, all three GWA studies use a partially overlapping set of samples, namely, the AGRE samples. Still, all of the reported associations are at distinct loci, even the two that are located on the same chromosome 5p region, and evidence for the reciprocal best loci could not be detected. Since the availability of ASD samples in the field is limited, the publicly available AGRE sample collection is in wide use, introducing challenges to the interpretation of results in a wide variety of genetic studies which use partially (or completely) overlapping study samples. Yet, despite the questions that arise from the lack of overlap in the results of these studies, they nevertheless demonstrate that effect sizes for common variants in ASDs are very low (odds ratios for the best SNPs in all three studies range from 0.53 to 1.19) and massive study samples are required to identify these variants reliably. Taken together, these studies strongly suggest that common variants explain only a small fraction of the genetic background of ASDs, and alternative methods and larger study samples are required to address the role of rare genetic events. However, given the heterogeneity of the disorder, it is likely that most of the predisposing factors turn out to be family or subgroup specific and cannot be found simply by increasing sample size.

2.2.7 Structural variation

Before CNV analysis was made possible by high-throughput SNP arrays and the study of chromosomal abnormalities relied on traditional cytogenetic methods, it was estimated that cytogenetic abnormalities occured in ASDs with a frequency of 3-5% (Ritvo *et al.* 1990, Chakrabarti and Fombonne 2001, Wassink *et al.* 2001a, Wassink *et al.* 2001b, Reddy 2005). Based on the information obtained from recent CNV studies, the current frequency estimate is 6-7% (Marshall *et al.* 2008) and this number is likely to increase as resolution in CNV detection improves. Also, including individuals with dysmorphic features and intellectual disability increases the estimate.

Like linkage, cytogenetic abnormalities in ASDs have been described in almost all chromosomes. The most frequent finding is to chromosome 15q11-q13 (Gillberg 1998), where especially inherited duplications are common and account for 1-2% of all aberrations in ASDs (Freitag 2007, Abrahams and Geschwind 2008). Deletions at this locus are known to cause two cytogenetic imprinting disorders, Angelman syndrome (MIM#105830) and Prader-Willi (MIM#176270) syndrome, depending on whether the deleted region is of maternal or paternal origin. In ASDs, the duplications are usually of maternal origin. Yet, autistic features appear to be present in both Angelman syndrome and PWS.

The first systematic genome-wide search for DNA copy number variation (CNVs) in autism was published in 2007 (Sebat *et al.* 2007). The study focused on idiopathic autism and identified an excess of *de novo* CNVs, i.e. CNVs not present in the respective parents, in individuals with autism compared with controls using comparative genomic hybridization (CGH). Validated *de novo* CNVs were identified in 12 out of 118 (10%) of patients with sporadic autism, in 2 out of 77 (3%) of patients with an affected first-degree relative, and in 2 out of 196 (1%) of controls. In addition to identifying individually interesting CNVs and thereby providing clues of the possible pathology, this study reinforced the idea that there is a difference in the frequency of *de novo* CNVs between sporadic and familial autism cases, suggesting distinct mechanisms in each. This idea was later supported by a similar study (Marshall *et al.* 2008).

In conjunction with the GWA studies performed in ASDs, many of the large cohorts have now also been analyzed for CNVs. In a study by Glessner and colleagues (2009), CNVs were analyzed in the ACC cohort (859 cases and 1409 controls), the AGRE cohort (1336 cases), and additional controls (n=1110). Numerous candidate loci with an enrichment of CNVs were identified in ASD cases compared with controls, and the CNVs were found to target genes involved in neuronal cell adhesion and ubiquitin degradation. In the Autim Genome Project (AGP) study (Pinto *et al.* 2010), rare (<1%) CNVs were analyzed in 996 ASD cases and 1287 matched controls. Cases were found to carry a 1.19 fold higher global burden of rare CNVs, both *de novo* and inherited, especially at loci previously implicated in ASDs or intellectual disability. They also reported that the *de novo* CNV rate was roughly equal in simplex and multiplex families, which contrasts the previous studies. An interesting related observation is that the average age of fathers of affected children has increased (Reichenberg *et al.* 2006), which might partially explain the elevated *de novo* CNV rates and ASDs.

A few individual CNV findings have recently gained attention in particular. For example, a recurrent *de novo* deletion on chromosome 16p11 was identified by three independent autism studies (Kumar *et al.* 2008, Marshall *et al.* 2008, Weiss *et al.*

2008). The deletion spans ~500 kb, overlaps ~30 genes, and seems to be present in ~1% of autism cases in different cohorts. Occasionally, also the reciprocal duplication was observed in ASD cases. In one of the studies (Weiss *et al.* 2008), the deletion was found only from 2 out of 18 834 unscreened controls, which highlights the significant enrichment in individuals with autism. However, in the study by Glessner *et al.* (2009), CNVs at the 16p11 locus did not segregate to all cases within families and it was transmitted also to unaffected siblings, suggesting that the locus is not sufficient by itself to cause ASDs in these individuals. Overall, many of the identified CNVs overlap with previously implicated autism candidate genes, in particular genes encoding for synaptic cell adhesion molecules which further supports the role of synaptic dysfunction in ASDs (see Section 2.2.9).

To conclude, although the identification of rare CNVs in ASDs has generated a lot of excitement, it should be remembered that *de novo* CNVs are found also in the general population. Thus, distinguishing pathogenic CNVs from benign ones remains a challenge. Also, it is difficult to evaluate, whether a single CNV in a single affected individual is sufficient to cause the phenotype, or whether CNVs mediate their pathogenicity in conjunction with other CNVs or predisposing mutations or polymorphisms. To facilitate this task, resources such as the Autism Chromosome Rearrangement Database (http://projects.tcag.ca/autism), are finally emerging to tackle the increasing numbers of findings. Additional resources, such as the Decipher database (Firth *et al.* 2009) and the EUCARUCA database (European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations) provide phenotypic information of various microdeletion and duplication syndromes.

2.2.8 Gene expression studies

Genome-wide gene expression profiling in ASDs has greatly been hindered by the availability of samples, and only a handful of studies have been reported to date. Yet, despite being small, these studies have established that expression profiling, most commonly from lymphoblastoid cell lines (LCLs), can distinguish between affected individuals and their healthy siblings, as well as between different syndromic forms of autism (Baron *et al.* 2006a, Nishimura *et al.* 2007, Hu *et al.* 2009b), once again highlighting the importance of accurate phenotypic subgrouping in genetic studies. Overall, on the level of individual genes, less than 1% of the genes that have been identified as differentially expressed in ASD studies overlap between studies (Abrahams and Geschwind 2008). This is not surprising because autism is known to be heterogeneous, but also because the study samples are small, and gene expression levels are easily affected by various technical and environmental artifacts. However, among the overlapping genes are multiple genes from the imprinted region on chromosome 15q11-q13, including *UBE3A* (*ubiquitin*

protein ligase E3A), *NIPA2* (*non imprinted in Prader-Willi/Angelman syndrome 2*), and *CYFIP1* (*cytoplasmic FMR1 interacting protein 1*) (Abrahams and Geschwind 2008), possibly reflecting the common occurrence of cytogenetic abnormalities in ASDs at this locus (see Section 2.2.7).

To date (August 2010), altogether seven studies have performed genome-wide expression profiling of ASD cases and controls from LCLs, and one from brain-tissue (summarized in Table 4). A number of other studies have measured expression changes of individual, or a small number of, genes. Two of the genome-wide studies profiled syndromic forms of autism, i.e. autistic individuals with a known chromosomal aberration (Baron *et al.* 2006a, Nishimura *et al.* 2007). Expression profiling has also been performed in closely related phenotypes, such as Fragile X syndrome (Bittel *et al.* 2007) and chromosome 15q duplication cases (Baron *et al.* 2006b). The largest of these studies analyzed gene expression in LCLs from 116 cases and 29 controls obtained from the AGRE (Hu *et al.* 2009b). Using various ADI-R diagnostic scores, the authors applied multiple clustering algorithms on ~2000 individuals with autism (Hu and Steinberg 2009), and identified subgroups of autistic probands (116 in total) with different severity scores which were then used in global gene expression analysis. Fifteen genes regulating circadian rhythm were differentially expressed in the subgroup with most severely affected individuals, whereas 20 genes associated with androgen sensitivity were differentially expressed in all of the cases compared with controls. The authors suggest that this might underlie the 4:1 prevalence bias in affected males and females. Relatedly, it has been suggested that higher levels of testosterone in the developing fetus could produce behaviors relevant to those seen in autism. This theory is part of the "extreme male brain" theory of autism (Baron-Cohen 2002, Baron-Cohen *et al.* 2005, Knickmeyer and Baron-Cohen 2006) which has remained controversial.

Three studies have analyzed the genome-wide expression of micro-RNAs in ASDs, two in LCLs (Talebizadeh *et al.* 2008, Sarachana *et al.* 2010) and one in post-mortem cerebellum (Abu-Elneel *et al.* 2008). All three studies reported some putatively interesting differentially expressed miRNAs (ranging from nine to 43) and target genes for these miRNAs, which seem to relate to autism-relevant biological processes. However, the number of affected individuals in these studies is extremely small, ranging from five to thirteen. Two miRNA, hsa-miR-23a and 106b, were differentially expressed in two of these studies (Abu-Elneel *et al.* 2008, Sarachana *et al.* 2010), but the statistical methods of the other was proved invalid (Buyske 2009), thereby questioning the results. No studies specific to ASDs have addressed the role of specific miRNAs. However, there are few examples from related disorders, such as Rett (Nomura *et al.* 2008) and Tourette syndromes (Abelson *et al.* 2005).

In summary, gene expression studies in ASDs suffer from small sample sizes and limited statistical power. The number of studies will undoubtedly increase when more samples become available. While these studies have identified some putatively interesting differentially expressed genes and biological processes affected in ASDs, their biggest value has been in showing that gene expression profiles from LCLs can be used to distinguish different phenotypic subgroups in neuropsychiatric disorders.

Table 4. ***Summary of gene expression studies published to date (August 2010).*** *Only genome-wide studies are included. Studies focusing only on Fragile X syndrome or related disorders are not included.*

Reference	Phenotype	Samples	Tissue
Purcell *et al.* (2001)	Autism	10 cases, 23 controls	postmortem brain (mainly cerebellum)
Baron *et al.* (2006a)	Autism with dup(15q)	10 cases, 9 controls	LCL
Hu *et al.* (2006)	Autism, ASD	5 discordant MZ twin pairs and two non-autistic siblings, one control MZ twin pair	LCL
Melin *et al.* (2006)	Autism	6 cases, 6 controls	LCL
Nishimura *et al.* (2007)	Autism with FMR1-FM, autism with dup(15q)	15 cases (8 with FMR1-FM, 7 with dup15q), 15 controls	LCL
Gregg *et al.* (2008)	Autism, ASD	49 cases (35 AUT, 14 ASDs), 12 controls	Whole blood
Hu *et al.* (2009a)	Autism with severe language impairment	21 discordant sib-pairs (one affected, one non-autistic)	LCL
Hu *et al.* (2009b)	Idiopathic autism with three subgroups based on ADI-R severity scores	116 cases, 29 controls	LCL

ABBREVIATIONS: AUT=autism, ASD=autism spectrum disorder, LCL=lymphoblastoid cell line, MZ=monozygotic, FMR1-FM=Fragile X mutation, dup(15q)=chromosome 15q11-q13 duplication

2.2.9 Candidate gene studies

As with linkage studies, the number of candidate gene studies in ASDs is beyond the scope of this review. Therefore, only some of the key findings will be presented. Since Study I of this thesis focuses entirely on the *DISC1* gene, it will be separately reviewed in the following section.

Most of the positional candidate genes studied in ASDs fall between chromosome 7q22 and 7q36 regions, due to frequent linkage findings (Section 2.2.5). These genes include the *met proto-oncogene (MET)*, *reelin* (*RELN*), *contactin associated protein-like 2 (CNTNAP2)*, and *Engrailed 2 (EN2)*, amongst others. Of these, *RELN* is probably the most studied. Association of a 5' untranslated region (UTR) trinucleotide repeat polymorphism to ASDs has been reported by multiple genetic studies (Persico *et al.* 2001, Skaar *et al.* 2005, Serajee *et al.* 2006), and on the biological side, defects in reelin signalling in post-mortem cortices and reduced plasma levels of reelin have been detected in individuals with autism (Fatemi 2002, 2004).

Of the other genes at 7q, strong association and functional evidence was obtained for a common functional variant in the promoter of the *MET* gene, which disrupts *MET* transcription and cortical MET signalling in ASDs (Campbell *et al.* 2006, Campbell *et al.* 2007). The *EN2* gene has been linked to cerebellar abnormalities in mutant mice (Bauman and Kemper 2005), and association to autism has been detected in a few studies (Gharani *et al.* 2004, Benayed *et al.* 2005). However, one of the most convincing positional candidate genes in autism altogether is *CNTNAP2*. It is located in a linkage peak for a language-related QTL (Alarcon *et al.* 2002, Alarcon *et al.* 2005), and fine-mapping of the region revealed an association between a SNP in *CNTNAP2* and a language-related autism endophenotype. Subsequent studies have found evidence that both common and rare genetic variants, as well as cytogenetic and gene expression-related abnormalities in *CNTNAP2* increase the risk of ASDs (Alarcon *et al.* 2008, Arking *et al.* 2008, Bakkaloglu *et al.* 2008).

The imprinted locus on chromsome 15q11-q13 has been frequently studied due to the chromosomal abnormalities observed at this locus (see Section 2.2.7). The $GABA_A$ receptor subunit gene cluster, with *gamma-aminobutyric acid A receptor, beta 3 (GABRB3)* gene in particular, has received most of the attention, after GABA receptor density was found to be decreased in the hippocampus of individuals with autism (Blatt *et al.* 2001). Two other genes, *ATPase class V type 10C* (*ATP10C*) and the *ubiquitin-protein ligase E3A* (*UBE3A*) located in the maternal expression domain, are involved with the phenotypically similar Angelman syndrome.

However, mutations have not been identified at 15q and the overall association evidence remains inconclusive (Gupta and State 2007).

Synaptic cell-adhesion genes

The last few years have been exciting times in autism research, and the understanding of the molecular pathology in ASDs has taken a huge leap forward compared with what was known before. The identification of rare, high-penetrance mutations in genes encoding for various synaptic cell-adhesion molecules has directed the attention of researchers to the synapse, and, for the first time, a specific molecular mechanism explaining a piece of the puzzle is emerging. Although the identified mutations are rare and account only for a small fraction of all autism cases, related evidence is accumulating, and the hypothesis of synaptic dysfunction underlying ASDs has gained widespread acceptance (Garber 2007).

The first piece of evidence emerged, when eight females were reported to have *de novo* deletions of the short arm of the X chromosome (Xp22). Three of these females had autism (Thomas *et al.* 1999). Later, in an attempt to pinpoint the causative factors at the deleted region, mutations in two neuroligin genes (*NLGN3* and *NLGN4)* were identified and found to be associated with ASDs (Jamain *et al.* 2003, Laumonnier *et al.* 2004). Both mutations, a frameshift and an amino-acid changing substitution, were identified in two brothers with no other features of a genetic syndrome, and they arose *de novo* in the unaffected mothers. Both genes have subsequently been screened for mutations in several ASD cohorts, but so far, only rare events have been seen (Yan *et al.* 2005). Other members of the NLGN gene family have been less well characterized.

Neuroligins are cell-adhesion molecules that connect the pre- and postsynaptic membranes of glutamatergic and GABAergic synapses, and function in synaptogenesis during brain development (Song *et al.* 1999, Varoqueaux *et al.* 2004). Neuroligins bind another group of proteins at the synaptic junction, called neurexins, and together they are crucial proteins for trans-synaptic cell adhesion, and for aligning and activating synapses (Boucard *et al.* 2005). When mutations and deletions in *neurexin 1* (*NRXN1*) in individuals with autism were identified (Feng *et al.* 2006, Szatmari *et al.* 2007), substantial excitement was generated and the synaptic dysfunction hypothesis was fully adopted.

SHANK3 gene (*SH3 and multiple ankyrin repeat domains 3*), located at 22q13, was linked to ASDs and defects in synaptogenesis in a similar fashion. Different genetic lesions in *SHANK3* were identified in probands of three ASD families, including two deletions and a frameshift mutation. A later study screened ~400 individuals with ASDs and identified rare *de novo* variants in *SHANK3* in almost 1% of these cases (Moessner *et al.* 2007). *SHANK3* is a synaptic scaffolding protein which binds

both neuroligins and neurexins (Meyer *et al.* 2004). Recently, also *SHANK2* was implicated in ASDs and intellectual disability (ID). *De novo* CNVs were detected in the gene in two unrelated individuals with ASDs and ID, after which sequencing of ~600 affected individuals and controls revealed additional mutations in the gene (Berkel *et al.* 2010). Interestingly, also *CNTNAP2* belongs to the neurexin superfamily (Poliak *et al.* 1999).

Together with the recent evidence from autism GWA studies, which identified common variants in two cadherin genes, another group of cell-adhesion molecules (Wang *et al.* 2009b) (see Section 2.2.6), the recent large CNV scans in ASDs have further strengthened the role of the aforementioned genes by identifying rare CNVs which overlap with *NRXN1*, *NLGN1*, *SHANK2*, and *SHANK3* (Szatmari *et al.* 2007, Bucan *et al.* 2009, Glessner *et al.* 2009, Pinto *et al.* 2010). In summary, as evidence of synaptic dysfunction in autism continues to accumulate, this specific group of genes is by far the most interesting target for future genetic research in autism, and has the potential to contribute significantly to our understanding of the disorder. Yet, many of the other candidate genes studied in ASDs remain interesting as well, and warrant further studies. It is possible that when the biological processes affected in ASDs begin to unravel, unexpected connections between many of these genes are discovered.

2.2.10 *Disrupted-in-Schizophrenia-1 (DISC1)*

Genetic findings

DISC1 gene was originally identified in large Scottish pedigree, where a balanced translocation between chromosomes 1q42.1 and 11q14.3 was found to co-segregate with schizophrenia (SCZ) and related psychiatric conditions. The pedigree contains individuals with major depression, bipolar disorder (BPD), adolescent conduct disorder, anxiety, and minor depression (St Clair *et al.* 1990), suggesting that the translocation was not specific for SCZ but instead predisposes to an unspecific neuropsychiatric outcome. It was noticed that all translocation carriers, even the ones without a psychiatric diagnosis, had abnormal amplitudes of the P300 event-related potential, which is used as a measure for underlying cognitive defect (Blackwood *et al.* 2001). However, the protein coding gene disrupted by the translocation was named *Disrupted-in-Schizophrenia-1* and has been extensively studied as a candidate gene for schizophrenia since. However, it is now known that the neurobiological effects of *DISC1* are very wide-ranged, and the gene has been linked to many aspects of neurodevelopment and neuropsychiatric phenotypes.

The initial translocation finding was supported by genetic linkage observed at the *DISC1* locus at chromosome 1q42 in Finnish schizophrenia families (Ekelund *et al.*

2001, Ekelund *et al.* 2004). Later, specific SNP-haplotypes within *DISC1* were shown to be associated with SCZ and impaired visual working memory functions in SCZ (Hennah *et al.* 2003, Hennah *et al.* 2005). The most significant allelic haplotype was named "HEP3", and later studies have found association with either the same haplotype or SNPs in its vicinity in SCZ, BPD, schizoaffective disorder (Hodgkinson *et al.* 2004, Sachs *et al.* 2005, Thomson *et al.* 2005b, Maeda *et al.* 2006, Zhang *et al.* 2006, for e.g. DeRosse *et al.* 2007, Palo *et al.* 2007), and major depression (Hashimoto *et al.* 2006), as well as various endophenotypes defined using neurocognitive (Burdick *et al.* 2005, Thomson *et al.* 2005a, Liu *et al.* 2006) or imaging measurements (Callicott *et al.* 2005, Cannon *et al.* 2005). In some of these studies, the observed effects have been sex-specific, with most of the association signal originating from affected males. The role of *DISC1* in psychiatric illness has been extensively reviewed for example by Chubb and colleagues (2008).

The *DISC1* locus at 1q42 harbors two other genes as well. *DISC2 (Disrupted-in-Schizophrenia-2),* located antisense to *DISC1* and disrupted by the original translocation as well, is not known to encode a protein product. However, evidence exists that it might regulate *DISC1* expression through its non-coding RNA product (Millar *et al.* 2000b, Blackwood *et al.* 2001, Millar *et al.* 2001). *TSNAX* (*translin-associated factor X*) is located downstream of *DISC1* and remains less well characterized. Four intergenic exons are located between *DISC1* and *TSNAX* (see Section 4.3), and *TSNAX* has been observed to form fusion transcripts with *DISC1* through intergenic splicing involving these exons (Millar *et al.* 2000a).

Biology of DISC1

Based on current knowledge, DISC1 is considered a "hub" protein with extensive protein-protein interactions and a role in several pathways (Camargo *et al.* 2007). Disruption of these "DISC1 pathways", instead of *DISC1* only, has been proposed to mediate the risk for mental illness (Millar *et al.* 2003, Hennah *et al.* 2006), and association of the *DISC1*-interacting genes with schizophrenia and other related phenotypes has been observed (for e.g. Yamada *et al.* 2004, Burdick *et al.* 2008, Numata *et al.* 2009, Tomppo *et al.* 2009).

There is an increasing number of studies attempting to model *DISC1* disease biology and function in model organisms such as mouse, *Drosophila,* and zebrafish. Despite the obvious challenges related to phenotypic assessment of psychiatric disorders in other organisms, several disease-related traits and behavioural features can be studied, along with neuroanatomical and physiological features. Overall, numerous *Disc1* knockout and transgenic mice have been studied to date, most of which show behavioural and anatomical deficits that can be linked to psychiatric disease (Brandon *et al.* 2009). For example, mice of the 129S9 SvEv strain carry a deletion polymorphism in Disc1 exon 6, which introduces a premature stop codon in exon 7

(Koike *et al.* 2006). These mice display deficits in working memory, as reported also in schizophrenia (see previous section).

The information obtained from these studies, along with cellular studies suggest that the main functions of DISC1 relate to neurite outgrowth, neuronal migration, synaptogenesis, glutamatergic neurotransmission, microtubule network formation, and cAMP signalling (Chubb *et al.* 2008). Recently, DISC1 was shown to regulate neural progenitor proliferation by modulating canonical Wnt signalling via inhibition of GSKβ/β-catenin signaling (Mao *et al.* 2009). Most of the information of DISC1 function has come from the study of its binding partners (Figure 4).

For example, one of these binding partners, *phosphodiesterase 4B* (*PDE4B),* was shown to carry a balanced translocation in a subject diagnosed with schizophrenia and a relative with chronic psychiatric illness (Millar *et al.* 2005). *PDE4B* and *DISC1* were suggested to be interacting genetic factors in schizophrenia and to participate in the regulation of cAMP signaling. Interestingly, the same study also showed that *DISC1* and PDE4B expression level in patient lymphoblastoid cell lines carrying the original *DISC1* translocation t(1;11) or the *PDE4B* translocation t(1;16), respectively, was reduced with ~50%, suggesting that haploinsufficiency is the likely mechanism for SCZ susceptibility in these individuals.

Other well-known binding-partners of *DISC1* include for example *FEZ1* (*Fasciculation and elongation protein ζ-1*) and *NDEL1* (*nuclear distribution gene E homologue-like 1*). DISC1 and FEZ1 co-localize to the growth cones of cultured neurons (Miyoshi *et al.* 2003) and reduced *FEZ1* expression in rat hippocampal neurons was shown to result in neuronal defects affecting neuronal polarity, axon growth, and intracellular transport (Ikuta *et al.* 2007). Studies on the interaction of NDEL1 and DISC1 have elucidated the role of *DISC1* in the microtubule / centrosome cascade affecting neurite outgrowth (Ozeki *et al.* 2003). NDEL1 localizes to the centrosome along with other DISC1 binding proteins, and its knockdown has been shown to block neurite production, possibly due to deficient microtubule-based transport to the growing neurite tip (Kamiya *et al.* 2006).

In mouse, *DISC1* expression is highest at embryonic and postnatal stages which coincide with active neurogenesis and the onset of puberty (Schurov *et al.* 2004). It is expressed both in neurons and glia cells, predominantly in mitochondria with additional nuclear, cytoplasmic, and actin-associated locations evident (James *et al.* 2004). During embryonic development, *DISC1* is known to regulate neuronal migration and structural plasticity, and it has been shown that either the depletion of endogenous *DISC1* or the expression of mutated *DISC1* can impair neurite outgrowth and the proper development of the cerebral cortex *in vivo* (Kamiya *et al.* 2005). Thus, it seems that during cortical neurogenesis, *DISC1* acts to facilitate neuronal maturation. However, another study demonstrated that *DISC1* negatively

regulates neuronal maturation in the adult hippocampus (Duan *et al.* 2007), thereby suggesting a completely opposite role for *DISC1* in adult neurogenesis. Yet, defects in early cortical development are considered more important for example in schizophrenia (Ross *et al.* 2006), since neurogenesis is much more prevalent at that time compared with adulthood.

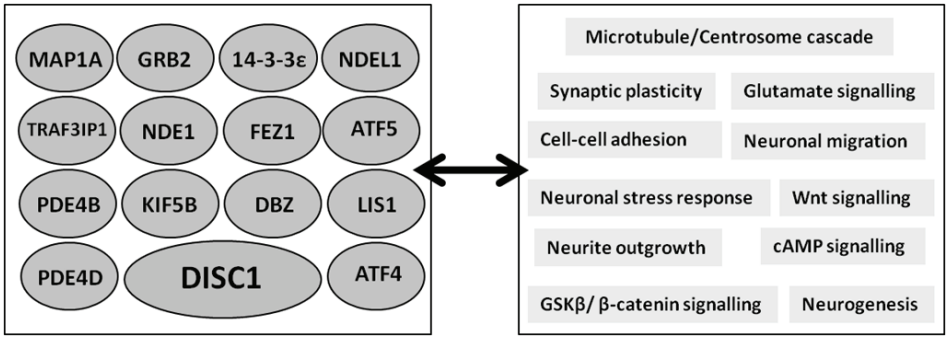


Figure 4. ***DISC1 interacting proteins and their functional role.*** *A selection of DISC1 binding and/or interacting proteins (left box) and related biological processes and cellular functions (right box) are displayed. Figure based on information from Chubb et al. (2008) and Brandon et al.(2009).*

It should be noted that *DISC1* has multiple isoforms which vary significantly in length, exon composition, and in the sequence of their 3'UTR sequences. The Uniprot database (www.uniprot.org) currently (August 2010) recognizes four different DISC1 protein isoforms produced by alternative splicing. These are named "long" (L), "long variant" (Lv), "short" (S), and "extra short" (Es), of which L has been chosen as the canonical sequence. However, for example the Ensembl database (www.ensembl.org) currently reports 11 alternative transcripts, of which ten produce a protein product, and RefSeq (http://www.ncbi.nlm.nih.gov/refseq) lists 23 alternative transcripts, including the four in Uniprot. Most of these variants lack multiple 3' exons and instead have alternate 3' segments compared with variant L. Further, other studies have demonstrated that the variability of *DISC1* splicing is extremely complex, especially in the brain, where more than 50 alternative transcripts were recently identified (Nakata *et al.* 2009). Thus, it is a huge challenge to relate the genetic and biological information available to the different isoforms, which are likely to reflect the subtle tissue-specificity and regulation of DISC1-mediated processes.

In conclusion, the genetic evidence for *DISC1* is consistent with the idea that multiple mechanisms, affected by several independent mutations, link pathways involving *DISC1* and its binding partners to clinical neuropsychiatric phenotypes. It has been suggested that such mutations can either i) cause reduced expression of all

DISC1 isoforms, ii) alter the spectrum of isoforms present, iii) alter the stability of DISC1, or iv) alter the ability of DISC1 to bind its interactors (Mackie *et al.* 2007). Current knowledge implies that aberrations in DISC1 function underlie a wide range of phenotypes, from severe psychiatric disorders such as schizophrenia to milder neuropsychiatric manifestations, such as Asperger syndrome. It remains to be seen whether also other members of the *DISC1* pathway are involved in these phenotypes, and whether the wide-range molecular effects of DISC1 turn out to be true also for other genes implicated in psychiatric illnesses.

3 AIMS OF THE STUDY

In this study we have taken two different approaches to study the genetic basis of autism spectrum disorders. The first is a candidate gene approach, in which we have focused on a single gene, *DISC1* (*Disrupted-in-schizophrenia-1*), and performed targeted experiments to address a highly specific question. The second is a hypothesis-generating genome-wide approach, in which we have used different levels of genomic data to gain information of the genetic basis of autism spectrum disorders in a group of genealogically connected individuals from a population sub-isolate, and started to explore the underlying biological processes in these individuals.

More specifically we have aimed to:

I. Study whether genetic polymorphisms in *DISC1* predispose to autism spectrum disorders (Study I)

II. Investigate *in vitro* whether genetic polymorphisms in *DISC1* affect the regulation of its expression by specific micro-RNAs in an allele-specific manner (Study I, unpublished data)

III. Comprehensively characterize the genetic architecture of autism spectrum disorders in genealogically connected individuals from Central Finland in order to identify the predisposing genetic variants, shared haplotypes, and biological pathways (Studies II and III).

4 MATERIALS AND METHODS

All of the methods used in this study are described in detail in the original publications. An overview of the methods is given below, with an emphasis on the methods used with the unpublished data.

4.1 Study sample

The Finnish families used in this study were recruited by the National Institute for Health and Welfare (previously called the National Public Health Institute) in Helsinki through university and central hospitals in collaboration with the University of Helsinki. All autism spectrum disorder diagnoses were assessed according to detailed structured interviews based on the diagnostic criteria in the International Classification of Diseases, 10th Revision (ICD-10) (World Health Organization 1993) and the Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV) (American Psychiatric Association 1994). Supplementing diagnostic information was collected with instruments such as the Childhood Autism Rating Scale (CARS), Asperger Syndrome Screening Questionnaire (ASSQ) (Gillberg and Gillberg 1989, Ehlers and Gillberg 1993, Ehlers *et al.* 1999), Asperger Syndrome Diagnostic Interview (ASDI) (Gillberg *et al.* 2001) and criteria proposed by Gillberg and co-workers (Gillberg and Gillberg 1989, Ehlers and Gillberg 1993). All of the study subjects underwent throrough medical and clinical examinations, including neurological examinations and psychological and neuropsychological evaluations. Diagnoses were ascertained by multidisciplinary teams with extensive experience and common training, using the same set of diagnostic instruments. The autism diagnoses have later been validated using the Autism Diagnostic Interview - Revised (ADI-R) questionnaire (Lord *et al.* 1994), the gold standard of the field, which resulted in 96% consistency with the original diagnoses (Lampi *et al.* 2010). This study has been approved by relevant ethical committees, and informed written consent was received from all of the participating families. The different study samples are presented below, and exact numbers of families, individual cases, and controls analyzed in Studies I-III are listed in Table 5.

4.1.1 Autism families

The autism study sample consists of 97 Finnish families with altogether 138 affected individuals (105 males, 33 females) diagnosed with childhood autism (AD; autistic disorder), Asperger syndrome (AS) or pervasive developmental disorder not otherwise specified (PDD-NOS). Only families with at least one child with autism

are included. Families with associated medical conditions such as Fragile X syndrome or profound intellectual disability were excluded. Individuals with AS and PDD-NOS were included in this study sample because approximately one-third of the probands with autism had a first-degree relative with these conditions. The affected individuals are divided into three liability classes based on their diagnosis. Individuals with strictly defined autism comprise the first liability class (LC1), whilst LC2 includes also individuals with AS, and LC3 all affected individuals (autism, AS, PDD-NOS).

4.1.2 Asperger syndrome families

The Asperger syndrome (AS) study sample consists of 29 large Finnish pedigrees with only individuals with AS in multiple subsequent generations. The sample does not overlap with the autism study sample. The total number of affected individuals is 143 (85 males, 58 females) of which 119 completely fulfill the ICD-10 criteria for AS and comprise the liability class 1 (LC1). Additional 24 have AS-like features but do not completely meet all of the diagnostic criteria (LC2). Only individuals with normal cognitive development before the age of three were included.

4.1.3 Extended ASD pedigrees originating from Central Finland

A central part of this thesis is formed by the analysis of a large, extended ASD pedigree from Central Finland (CF). This pedigree (subsequently referred to as Pedigree 1) consists of 18 Finnish families (20 nuclear families), which have been genealogically traced back to the 17th century and found to originate from common ancestors (Figure 5). Altogether, the pedigree has 34 affected individuals (25 males, 9 females), of which 17 are diagnosed with autism (including a monozygotic twin pair, of which only one used in analyses), 14 with AS, and three with PDD-NOS. Of the families in the pedigree, 14 are included in the autism sample, one in both autism and AS sample, and three in neither. The pedigree was first described by Auranen and colleagues (2003) after which six new families have been added and more detailed genealogical links established, as presented in Study II.

In Study III, we also analyzed a second, smaller extended autism pedigree originating from the same CF region (Pedigree 2) (Figure 6). This pedigree consists of eight families with 11 affected individuals (seven males, four females), of which nine are diagnosed with autism, one with AS, and one with PDD-NOS. The affected individuals from Pedigree 1 and 2 were combined to form the CF-GWAS case-control dataset, which, in addition to the 27 cases from the two pedigrees (one affected individual per nuclear family), contained 24 individuals with at least two

grandparents born in the same Central Finland region, but with no genealogical links to the pedigrees.

All genealogical studies were carried out using local church and civil registers (for information post year 1850) and the Finnish National Archives for the earlier periods, in accordance with published criteria (Varilo 1999).

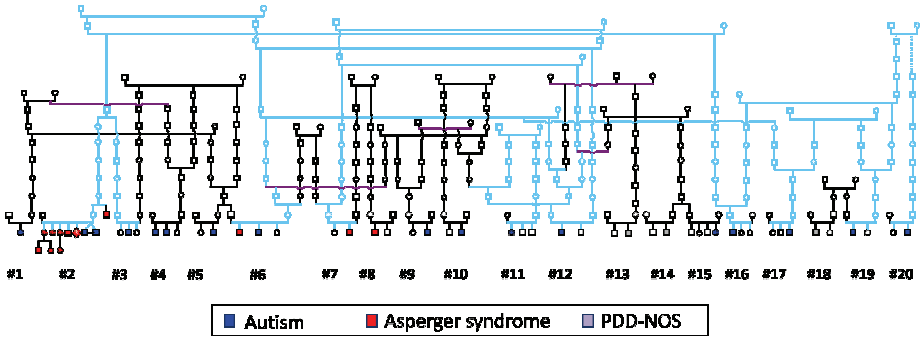


Figure 5. ***Pedigree 1 originating from Central Finland.*** *The core pedigree and links to the common ancestors are marked in blue. Abbreviations: PDD-NOS=pervasive developmental disorder not otherwise specified.*

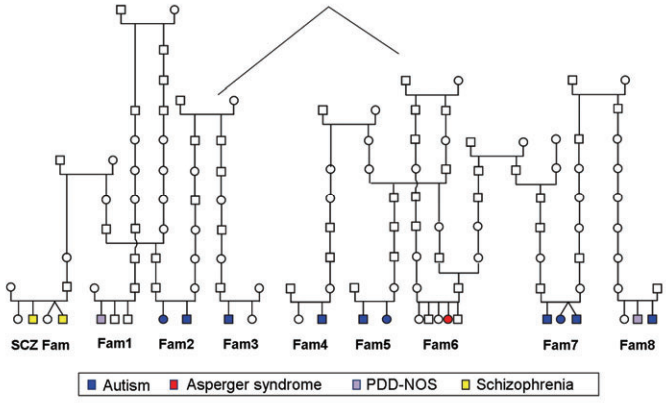


Figure 6. ***Pedigree 2 originating from Central Finland.*** *The schizophrenia family was not used in this study. Abbreviations: Fam=family, SCZ=schizophrenia, PDD-NOS=pervasive developmental disorder not otherwise specified.*

4.1.4 Control samples

In Studies I and II, all analyses involving Pedigree 1 were performed using independent regional controls (n=93) in order to properly control for the diversity of background alleles in different regional study samples. The controls were collected from the same geographical area where the families in the extended pedigree originate. In Study I, randomly selected trios representative of the Finnish population (n=60) were additionally used as population controls to enable allele-frequency comparisons with the other *DISC1* association studies performed in Finnish study samples.

In Study III, controls samples for the CF-GWAS dataset (n=181) were matched from a large available pool of Finnish controls with GWAS data available using complete linkage agglomerative clustering based on pairwise identity-by-state (IBS) distance. In Study III, we additionally used three publicly available datasets as controls in the pathway analysis. The autism GWAS dataset consists of 1001 families with 1529 individuals with autism from the US population (Wang *et al.* 2009b), and was obtained from the Autism Genetic Resource Exchange (AGRE; http://www.agre.org) (Geschwind *et al.* 2001). The autism gene expression dataset (Hu *et al.* 2009b) was obtained from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/, dataset ID: GSE15402). It consists of 87 males with idiopathic autism and 29 non-autistic, roughly age-matched controls, also from the AGRE. In order to reduce genetic heterogeneity and to increase comparability between the publicly available autism datasets, we re-analyzed the AGRE-GWAS dataset after excluding all siblings with milder ASD phenotypes. The third dataset is a Crohn's disease (CD) GWAS dataset from the Wellcome Trust Case Control Consortium (WTCCC) (2007), consisting of 1748 cases and 2938 controls. The two autism datasets are amongst the largest autism studies published to date, whilst the Crohn's disease dataset represents a phenotype with a relatively well established biology for which certain pathways can be expected to be found.

Table 5. ***Description of samples analyzed in each study.*** *The total number of individuals denotes the number of individuals, for which a DNA sample was available in each study.*

Study sample	Families	Individuals	Affected individuals						Study
			Total	Males	Females	LC1	LC2	LC3	
Autism families	97	356	138	105	33	118	8	12	I, II
AS families	29	210	143	85	58	119	24	x	I, II
Autism+AS families combined	126	566	218	190	91	237	32	12	I
CF Pedigree 1	18(20)	34	33*	24	9	16(17)	14	3	I, II, III
CF Pedigree 2	8	11	11	8	3	9	1	1	III
CF-GWAS (includes Pedigrees 1 and 2)	x	51	51	39	12	45	4	2	III
CF regional controls 1	x	93	x	NA	NA	x	x	x	I, II
CF regional controls 2[a]	x	181	x	115	66	x	x	x	III
CF-EXPR cases	x	10	10	10	x	9	x	1	III
CF-EXPR controls	x	10	x	10	x	x	x	x	III
Finnish control trios	60	180	x	106	74	x	x	x	I
AGRE-GWAS[b]	1001	2460	1500	1203	297	1500	x	x	III
AGRE-EXPR cases	x	87	87	87	x	87	x	x	III
AGRE-EXPR controls	x	29	x	29	x	x	x	x	III
CD-GWAS	x	1748	1748	682	1066	NA	NA	NA	III
CD-GWAS controls	x	2938	2938	1440	1458	x	x	x	III

ABBREVIATIONS: AS=Asperger syndrome, CF=Central Finland, EXPR=global gene expression study, GWAS=genome-wide association study, LC=liability class, CD=Crohn's disease, AGRE=Autism Genetic Resource Exchange, NA=not applicable

[a] Dataset overlaps partially with CF regional controls 1.

[b] Dataset pruned to include only families with at least one LC1 autism case. Additional siblings with LC2 or LC3 phenotype were also excluded.

4.2 Genotyping

Genomic DNA was extracted from EDTA-treated peripheral blood samples using the Puregene DNA purification system (Qiagen) according to manufacturer's instructions, or by using a phenol-chloroform protocol modified from the original protocol by Vandenplas and colleagues (1984). Most of the DNA samples were extracted and processed by the DNA Extraction Unit of the National Institute for Health and Welfare (Helsinki).

Genotyping of single-nucleotide polymorphisms (SNPs) in Studies I and II (n=11 and n=152, respectively) was primarily done using the homogenous MassEXTEND (hME) and iPLEX technologies of the Sequenom MassARRAY platform (Sequenom Inc.), as specified by manufacturer's instructions. In Study II, some of the SNPs were additionally genotyped using fluorogenic 5' nuclease allelic discrimination chemistry (TaqMan) with the ABI Prism 7900 Sequence Detection System (Applied Biosystems).

Genotyping of the microsatellite markers in Study I (n=2) was done using the ABI 3730 DNA sequencer (Applied Biosystems). In Study II, microsatellites of the initial genome-wide scan (n=1109) were genotyped by deCODE Genetics Inc. (Reykjavik, Iceland), whereas the follow-up microsatellites (n=44) were genotyped as in Study I. The average intermarker distance in the genome-wide scan of Study II was 3.43 cM.

In Studies I and II, all genotypes were checked for correct Mendelian transmission using PEDCHECK v.1.1 software (O'Connell and Weeks 1998) and monitored for Hardy-Weinberg equilibrium (HWE) and duplicate sample accuracy. All markers accepted for analysis displayed a minimum genotyping success rate of 90%, with the majority of markers having a success rate of over 95%. The borderline for the minor allele frequency (MAF) of SNP markers was 5%, with most of the SNPs having a MAF over 10%.

In Study III, the genome-wide SNP data for the CF-GWAS study sample was produced at the Broad Institute (Boston, USA) using the Illumina HumanHap 300 Beadchip, which includes over 317 000 tag SNPs derived from the Phase I of the International HapMap project (2005). The regional controls were genotyped with the Illumina HumanHap 370 and 550 Beadchips at the Broad Institute and at the Wellcome Trust Sanger Institute (Cambridge, UK), which include over 317 000 and 550 000 tag SNPs from HapMap Phase I+II, respectively. The publicly available AGRE-GWAS dataset has been genotyped with Illumina HumanHap 550 Beadchip, and the CD-GWAS with Affymetrix GeneChip 500K arrays, both of which include over half a million of SNPs.

All quality checks of the genome-wide SNP data were performed using PLINK software (Purcell *et al.* 2007). Genotypes were checked for correct Mendelian transmission when family information was available. To exclude sample swaps and contamination, gender checks were performed, and the samples monitored for identical-by-descent (IBD) sharing and mean heterozygosity. For GWA analysis, we only used SNPs genotyped in both case and control datasets. The following basic quality control criteria were applied: genotyping calling rate per SNP and per sample > 90%, MAF > 5%, HWE > 0.0001.

4.3 Study of *DISC1* as a candidate gene for ASDs (Study I and unpublished data)

In Study I, we carried out a targeted replication of a previous *DISC1* association study performed in Finnish families ascertained for schizophrenia (Hennah *et al.* 2003) to investigate the role of *DISC1* in ASDs. We analyzed the previously reported markers and haplotypes both separately and jointly in 97 Finnish autism families and 29 AS families, as well as in Pedigree 1 originating from Central Finland. Further, the observed association (see Section 5.1.1) led us to search for explanations to the wide-ranging neurobiological effects observed with DISC1. Consequently, we identified polymorphic miRNA target sites in the gene and decided to study them further with the hypothesis that altered miRNA regulation of *DISC1* would provide a plausible explanation for the wide-range of disease associations and neurobiological effects observed with the gene.

4.3.1 Association and haplotype analysis

We analyzed altogether 11 SNPs and two microsatellite markers spanning a ~600 kb region on chromosome 1q42 with *DISC1*, *DISC2*, and *TSNAX* genes (see Section 5.1). Since the aim of this study was to replicate previous findings for *DISC1* rather than fully analyze the genetic variation in the gene, markers were chosen based on previous *DISC1* studies in schizophrenia and bipolar disorder. The two microsatellite markers were chosen according to previous linkage studies in the Finnish population (Ekelund *et al.* 2001, Ekelund *et al.* 2004) and the SNPs according to the original Finnish *DISC1* association study (Hennah *et al.* 2003) and a later study by Thomson and colleagues (2005b).

All analyses were performed separately for the autism and AS study samples, as well as CF Pedigree 1. In addition, autism and AS families were analyzed jointly as a broad ASD phenotype (Table 5). Due to previous evidence of sex-dependent

effects with *DISC1* and the overall higher prevalence of ASDs in males, we performed statistical analyses also with affected males only (autism sample n=105, AS sample n=85), using the genotypes of females only for phase determination. In all statistical analyses of this thesis, the ASD phenotype was analyzed as a binary trait, that is, all individuals were treated either as affected or unknown, which is common practice especially in most complex disease studies where it is often challenging to assign individuals as healthy.

Both single marker and haplotype association analyses were performed using FBAT 1.5.5 (Horvath *et al.* 2001) and TRANSMIT 2.5.4 (Clayton 1999) software. Both are family-based transmission disequilibrium tests (TDT) for alleles and haplotypes, and are able to test for transmission even with incomplete parental genotype data and unknown phase. In all analyses, we accounted for possible effects of linkage on the results by using the empirical variance option of FBAT and performing 100 000 bootstrap replicates in TRANSMIT.

Haplotype analysis was restricted to four specific haplotypes (named "HEP1-4") associated with SCZ in Finnish families (Hennah *et al.* 2003) in order to reduce the multiple testing burden. We tested the exact same haplotypes as in the original study, except in the case of HEP2 where rs1630250 and rs1655285 were used as surrogates for the haplotype information provided by the original haplotype consisting of rs1615344, rs1615409 and rs766288 (see Section 5.1).

Pseudomarker vs. 0.9.7 beta program (Göring and Terwilliger 2000) was used as the sole method of analysis with Pedigree 1, due to its capability to handle complex pedigree structures. Since Pedigree 1 has only a small number of informative transmissions available, FBAT and TRANSMIT were not applicable. With Pseudomarker, single marker association and linkage in different kinds of pedigree structures can be analyzed both separately and jointly. It is also capable of incorporating data from independent controls to a family-based analysis to improve the estimation of allele frequencies, thus enabling us to use the regional controls in the same analysis with Pedigree 1. Haplotype association analysis was not performed in Pedigree 1. Pseudomarker was also used to monitor for two-point linkage in all study samples.

4.3.2 Polymorphic miRNA target site prediction

We used two available databases to search for polymorphic miRNA target sites in *DISC1*, Patrocles (http://www.patrocles.org) (Hiard *et al.* 2010) and PolymiRTS (http://compbio.utmem.edu/miRSNP/home.php) (Bao *et al.* 2007). Since the aim was to look specifically for polymorphic miRNA target sites, we did not address the multitude of non-polymorphic miRNA target sites predicted for *DISC1*.

To predict the target sites, PolymiRTS uses same criteria as TargetScan, one of the most commonly used miRNA target prediction algorithms (Lewis *et al.* 2005, Grimson *et al.* 2007, Friedman *et al.* 2009). Perfect Watson-Crick base-pair match is required for the target seed nucleotides 2-7. Additionally, either a perfect match for seed nucleotide 8 or an "anchor" adenosine immediately downstream of the 2-7 seed in the target is required. PolymiRTS looks only for SNPs located in the 3'UTR regions of genes and requires the predicted target site to be present in at least two other vertebrate genomes. It is also capable of categorizing the identified target sites based on "wobble" pairing, which refers to situations where A/G alleles in the target mRNA are able to form G:U wobble base pairs with the miRNA, leading to a possibly less deleterious effect.

Patrocles also restricts its search to SNPs in the 3'UTR, and requires the target site to be present in at least one other primate genome. It uses two separate sets of target motifs. The so called "X-motifs" are a group of 540 octamers identified on the basis of unusually high motif conservation scores in the 3'UTR (Xie *et al.* 2005). "L-motifs" meet the TargetScan criteria, as described above.

In the beginning of this study, only the four most well-known isoforms of *DISC1* were properly annotated and known (see Section 2.2.10). Therefore, all the polymorphic target predictions take into account only these transcripts (L, Lv, S, Es). For the purpose of this study, we limited our search only to the two longest isoforms, L and Lv.

4.3.3 *DISC1* expression constructs

In order to functionally validate the identified polymorphic miRNA target sites, we created expression constructs of *DISC1*. A full length EST clone of *DISC1* Lv (long variant isoform) including full 3'UTR and 5'UTR sequences was obtained from the IMAGE consortium (Lennon *et al.* 1996) (clone ID: IMAGE:9007180). An overexpression construct was created by cloning the full length gene (6856 bp) into a pcDNA3.1(+)/Hygro vector (Invitrogen) by creating suitable artificial restriction enzyme (RE) sites using short oligoduplexes harbouring the desired site (Figure 7, Table 6). The presence and correct orientation of the insert was verified by restriction analysis and direct sequencing. Expression constructs harbouring the alternative SNP alleles were created using the QuikChange Lightning Site-Directed Mutagenesis Kit according to manufacturer's instructions (Stratagene). Mutagenic primers were designed using QuikChange Primer Design Program (http://www.stratagene.com/ qcprimerdesign) and are provided in Table 7.

Table 6. ***Properties of the synthetic oligoduplex used in DISC1 Lv cloning.*** *The core structure denotes the restriction enzyme sites that were introduced into the plasmid.*

Core structure	Sequence	Length (bp)	Tm	Overhang
5’-NheI-X-SalI-X-BssH II-X-KpnI-3’	5'-CTAGCAGTCGACATTAGCGCGCAGGTAC-3'	28	66°C	NheI, KpnI
3’-X-SalI-X-BssH II-X-5’	3'-GTCAGCTGTAATCGCGCGTC-5'	20	61°C	none

ABBREVIATIONS: Tm=melting temperature, bp=base pair

Table 7. ***Mutagenic primers used in site-directed mutagenesis.*** *Primers designed with QuikChange Primer Design Program (Stratagene).*

#rs	wt allele	Mutant allele	Primer sequence (5’ to 3’)	Length (bp)	Tm
rs11122396	A	G	gtcatgttttaaccacaagccgtaactcatctgttgtctttgc	43	79.7°C
			gcaaagacaacagatgagttacggcttgtggttaaaacatgac	43	79.7°C
rs980989	G	T	ttgccatgctaagccctttacattcatatcctataatccc	40	78.5°C
			gggattataggatatgaatgtaaagggcttagcatggcaa	40	78.5°C
rs9308481	G	A	tctaaggcacagagctggtaaaatatgaagtaatagtgaacc	42	78.7°C
			ggttcactattacttcatattttaccagctctgtgccttaga	42	78.7°C
rs11803088	C	T	atgccttcttgtatatgtaattcaatacttttatttttaatatcctcaccttatctaat	59	78.1°C
			attagataaggtgaggatattaaaaataaaagtattgaattacatatacaagaaggcat	59	78.1°C

ABBREVIATIONS: wt=wild type, bp=base pair, Tm=melting temperature

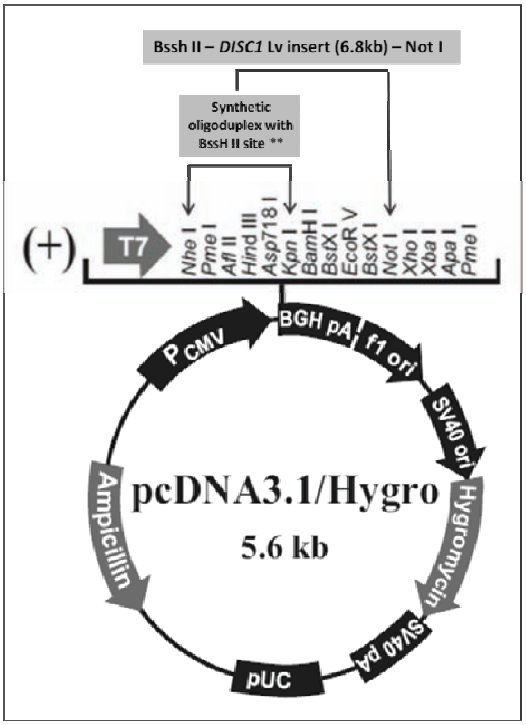


Figure 7. ***DISC1 Lv overexpression construct.*** *The cloning sites for the synthetic oligoduplex and the DISC1 Lv insert are shown. Vector map provided by Invitrogen.*

4.3.4 Cell culture and transfections

In order to study, whether the miRNAs predicted by Patrocles and PolymiRTS would have an allele-specific effect on *DISC1* expression in HEK293FT cells, we designed a set of different transfection experiments. The overall experimental setup is summarized in Figure 8.

The experiment was divided into three parts. First, we measured effects of the selected miRNAs on endogenous *DISC1* expression in 293FT cells. Second, we measured the effects of the same miRNAs in 293FT cells transiently overexpressing the wild type *DISC1* Lv construct. Thirdly, we repeated the overexpression experiment using constructs with the alternative SNP alleles with the best miRNAs from the first two rounds. We used Pre-miR miRNA precursor molecules (Ambion) to mimic and enhance the effect of the endogenous miRNA. In addition to the miRNAs tested, each transfection experiment contained untransfected cells, one or two siRNAs against *DISC1* (Sigma MISSION® siRNA), and a Pre-miR precursor negative control miRNA (Ambion) to monitor for unspecific technical artefacts. The

negative control miRNA was used as the baseline for monitoring miRNA-induced changes in expression. Each transfection experiment contained three to four biological replicates, of which three with the most consistent RNA yield were selected for qPCR. Each experiment was independently replicated three times.

293FT cells (Invitrogen) were cultured in Dulbecco's Modified Eagle Medium (D-MEM) (Gibco) with L-glutamine, 4500 mg/L D-Glucose, and 110 mg/L sodium pyruvate. The medium was supplied with 10% FBS, 1% penicillin, and 1% streptomycin, and cultured in +37°C and 5% CO2.

Transfections were performed in 96-well format using 6000 cells per well, miRNA and siRNA concentration of 30 nM per reaction, and in co-transfections 100ng of plasmid DNA per reaction, as recommended by the manufacturer. We performed transfections using the siPORT NeoFX reverse transfection protocol (Ambion). In this protocol, the transfection complexes are pipetted to the plate and overlaid with the desired number of cells in a fixed volume. Briefly, cells were counted and diluted suitably to seed 6000 cells per well in 80 µl volume. The transfection reagent, Pre-miR miRNA precursor molecules, siRNA molecules, and plasmid DNA (in the overexpression experiments) were diluted in plain D-MEM, and the transfection complexes prepared according to manufacturer's instructions. Transfection complexes (20 µl) were dispensed to the 96-well plates and overlaid with the cell suspension (80 µl). Cells were assayed after 48 h.

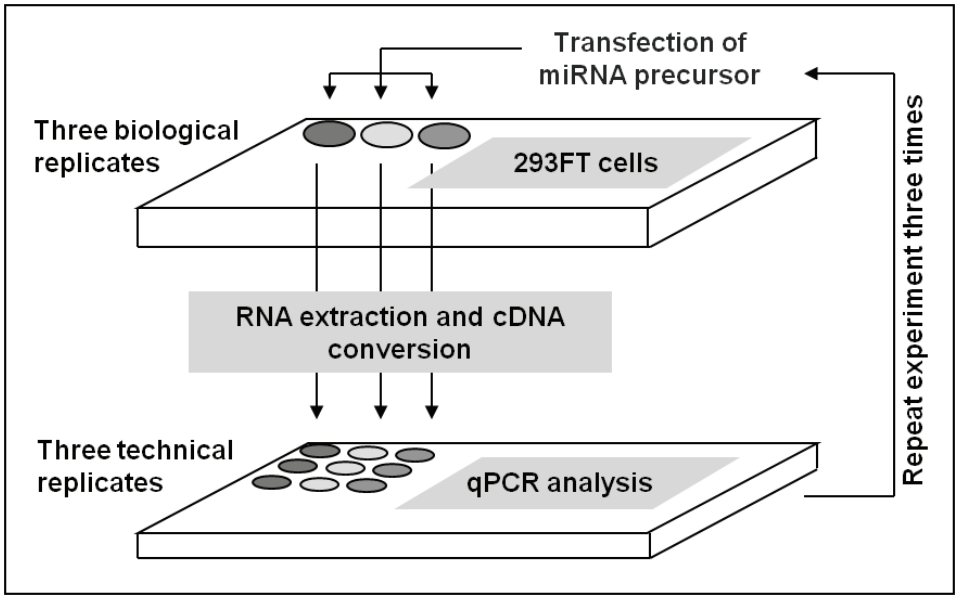


Figure 8. ***Overview of the experimental setup.***

4.3.5 RNA extraction and quantitative PCR

Total RNA from transfected cells was extracted using TRI Reagent (Molecular Research Center, Inc.). Cells were lysed with 120 µl of TRI Reagent and total RNA extracted according to manufacturer's instructions. We supplemented the protocol with manual Phase Lock Gel tubes (5Prime) to facilitate the separation of the organic and aqueous phases during extraction. RNA was treated with DNaseI enzyme (Fermentas) to remove any residual genomic DNA, and converted to cDNA using High Capacity RNA-to-cDNA Master Mix (Applied Biosystems) according to manufacturers' instructions. Quantitative Real-Time PCR was performed with 7900HT instrument using SYBR Green PCR Master Mix (Applied Biosystems) and the absolute quantification method. A five-point standard curve for each primer pair was included in each run, and absolute expression values for each sample were extracted based on the standard curves. Human *GAPDH* gene was used as the reference gene. *DISC1* expression was monitored with one pair of primers, picked from a test of eight primer pairs, designed to the boundary of exons 7 and 8 of *DISC1* with Primer Express 2.0 (Applied Biosystems) to avoid signal from residual genomic DNA (primer sequences provided in Table 8). Since exons 7 and 8 are only present in the longer *DISC1* isoforms (L and Lv), the qPCR signal should capture only the expression of these isoforms. Raw qPCR data was analyzed using the SDS 2.3 software (Applied Biosystems).

More specifically, qPCR was performed in 10 µl reaction volume in 384-well format. Each reaction contained a total of 10 ng of cDNA template, 5 µl of SYBR-mix, and 0.7 pmol of each primer. The standard curve samples had a known amount of 0.2, 2, 10, 20, and 50 ng per reaction. Each qPCR experiment (done on 384-well plates) contained samples from one transfection experiment in three technical replicates. Since each transfection experiment contained three biological replicates, and was additionally repeated three times, we had in total nine replicates of each test.

Table 8. ***Quantitative PCR primers sequences for human DISC1 and GAPDH*.**

Primer	Sequence	Length (bp)	Tm	GC
GAPDH p1	AACAGCGACACCCATCCTC	19	60	58
GAPDH p2	CATACCAGGAAATGAGCTTGACAA	24	62	46
DISC1 p1	AAAATCCCTCAACTTGTCACTTAAAGAA	28	60	32
DISC1 p2	CTCAGGGTGCTGCAGAATTTC	21	59	52

ABBREVIATIONS: Tm=melting temperature, GC=GC content of the sequence, bp=base pair, p1=forward primer, p2=reverse primer

4.3.6 Statistical analysis of qPCR data

Careful quality control was carried out to the qPCR data in order to ensure high data quality and avoid sources of bias. For each sample, we required the Ct values of all three technical replicates to fall within one Ct value. PCR efficiency was estimated from the slope of the standard curve, and the correlation coefficient of the curve monitored for overall standard quality. A melting curve analysis was performed at the end of each reaction to verify specific PCR amplification. Relative expression values of *DISC1* for all samples and replicates independently were obtained by dividing the absolute expression value of *DISC1* with that of *GAPDH*. A mean relative quantity was calculated to represent each biological replicate from the three technical replicate values. This value was then used in the subsequent statistical analyses.

The statistical significance of the effects of the miRNAs (and siRNAs) was assessed using a linear regression model with *DISC1* expression level as the dependent (linear) variable. Linear regression was calculated between the mean relative quantity of *DISC1* and the outcome, separately for each miRNA or siRNA and the negative control (altogether nine measurements from three different experiments). The negative control miRNA was used as the baseline in all comparisons. To correct for possible plate effects we included different qPCR runs as covariates in the model. Given the initial small number of cells in each well, we wanted to be sure that random differences in RNA extraction efficiency or RNA yield were not influencing the expression values. Thus, we also included the original RNA concentration of samples as a covariate in the model. However, the RNA concentration did not have an impact on the results (data not shown) so it was eventually removed from the model. All statistical analyses were performed using functions in the R Stats packages in R-2.10.0 (http://www.r-project.org)

4.4 Genetic analyses in the Central Finland extended pedigrees (Studies II and III)

In Studies II and III we analyzed different types of genetic data from the two CF extended pedigrees and some additional individuals from the same region (CF-GWAS study sample). We analyzed Pedigree 1 for linkage and linkage disequilibrium using microsatellite markers and finemapped the most significant regions of linkage with SNP markers. The findings from Study II were followed up in Study III with a denser set of genome-wide SNP markers and additional gene expression and pathway analyses.

4.4.1 Genome-wide linkage and LD analyses in Pedigree 1 (II)

In the initial genome-wide scan, we analyzed altogether 1109 microsatellite markers using the decode Genetic Map (Kong *et al.* 2002). The fundamental hypothesis of the study was that the observed genealogical links in Pedigree 1 reflect identical-by-decent (IBD) sharing of the same ancestral susceptibility variant(s). Therefore, we wanted to maximally extract information of allele sharing both within and across the families in our analysis. Our primary approach was the joint analysis of LD and Linkage of single markers genome-wide implemented in the Pseudomarker vs 0.9.7 beta analysis program (Göring and Terwilliger 2000). Both dominant and recessive Pseudomarker analyses were conducted, although the structure of the pedigree and the obtained results support inheritance in a recessive-like fashion. The "dominant Pseudomarker analysis" is analogous to affected relative pair methods in large families, and it weights the sharing between parents affected with ASDs and their affected children more strongly than that between unaffected parents and affected children. In the "recessive Pseudomarker analysis" contributions of both parents are weighted equally. Genotypes from 22 regionally matched controls were included in the analyses to improve power in LD analyses as well as to better estimate allele frequencies in the linkage analysis.

To monitor for allele sharing within families, we additionally analyzed multipoint linkage in the pedigree using the non-parametric multipoint linkage (NPL) analysis of Simwalk2 v.2.91 software (Sobel and Lange 1996), which is especially suitable for complex pedigrees. Simwalk2 produces five NPL statistics, of which we chose to focus on two that have been shown to be best suitable for the analysis of dominant and recessive traits (Lange and Lange 2004). These are the "BLOCKS" and "MAX-TREE" statistics, which are referred to as "NPL_recessive" and "NPL_dominant" in this study, respectively.

Pedigree 1 includes one nuclear family which has altogether 13 individuals affected with an ASD (one MZ twin pair with autism and 11 individuals with AS). To avoid the dominance of this single family in the analyses, we ran all analyses with and without the 11 AS cases (designated as Set 2 and Set 1 analyses, respectively). One of the monozygotic twins was included in both sets. This strategy was employed throughout Study II.

4.4.2 Follow-up and candidate gene analysis in Pedigree 1 (II)

Based on the initial genome-wide scan with Pseudomarker and Simwalk2, we chose ten most significant loci for follow-up. We increased the density of markers at these loci by genotyping 44 additional microsatellites. Based on the follow-up results and previous evidence for involvement in ASDs for two of the loci, we chose three for

fine-mapping (1q23, 15q12, and 19p13). Regional candidate genes were chosen from these loci based on biological relevance to autism (see Section 5.2.2), and analyzed with SNP markers using Pseudomarker. The SNPs were chosen using the Tagger algorithm in Haploview program (http://www.broadinstitute.org/haploview) to maximize the capture of common variation in the investigated genes (Barrett *et al.* 2005, de Bakker *et al.* 2005). As in the initial scan, we included genotype data from regional controls in the analysis, and increased the number of controls to 93, in order to further increase the accuracy of allele frequency estimation in the analysis. Due to prior evidence in autism, two of these loci, 1q23 and 15q12, were analyzed also in the nationwide autism and AS study samples, in addition to Pedigree 1 to assess their possible significance outside the isolate.

4.4.3 Genome-wide SNP analyses (III)

After the microsatellite-based genome-wide study of Pedigree 1 was published (Kilpinen *et al.* 2009), a number of studies have shown that genetic heterogeneity in autism is substantial, and that rare, family specific variants and mutations are likely to explain the majority of autism cases (see Section 2.2.9). Thus, we wanted to make an effort to thoroughly dissect the genetic architecture of ASDs in Pedigree 1. We therefore increased the marker density by using genome-wide SNP data, and extended our original study sample of 18 families (Pedigree 1) with 33 additional families (including Pedigree 2) from the same CF subisolate (CF-GWAS) to follow up the original scan.

The two extended pedigrees were analyzed for shared regions of homozygosity (ROHs) to monitor for possible recessive susceptibility variants. ROHs exceeding 100 kb were identified and assessed for overlap among individuals. A minimum overlap of 50 kb was used as a cutoff and only ROHs shared between more than half of the affected individuals in each pedigree and homozygous for the same haplotype were included. Liberal cutoffs for both segment length and frequency were used so that no regions of interest would be missed.

To identify shared, enriched risk variants inherited in a dominant-like fashion, shared segment analysis was performed as described previously (Purcell *et al.* 2007). Since we assumed inheritance from a common ancestor, we required IBD sharing among more than half of the cases in each pedigree, and that the same alleles would be shared between all pairs of affected individuals in the region. Only non-correlated SNPs in linkage equilibrium were used (approximately 56000 SNPs), as regions with high LD result in false positive calls of IBD sharing. We primarily looked for extended segments shared IBD between pairs of individuals exceeding 1000 kb. To evaluate whether the obtained results were specific to the pedigrees, we

assessed all regions of interest from the homozygosity and shared segment analyses also in the 181 regionally matched controls.

Traditional genome-wide association analysis (allelic chi-square test with one degree of freedom) was performed with tools implemented in PLINK (Purcell *et al.* 2007). Since a case-control association study requires the analyzed cases and controls to be independent, i.e. not related, we only included one affected individual per nuclear family into the GWAS. Since the included cases nevertheless were distantly related, we performed IBD analyses to evaluate the degree of their relatedness. Since the cases were no more related to each other than to the controls, no corrections for relatedness were applied in the analysis. Aware of the insufficient statistical power of the dataset, GWAS was initially performed for the purpose of pathway analysis.

4.4.4 Analysis of differential gene expression in ASD cases and controls (III)

Genome-wide gene expression profiles in Study III were obtained from ten individuals with ASDs (nine with autism, one PDD-NOS) and ten controls matched for age (5-17 years) and sex (CF-EXPR). Profiles were produced from mononuclear lymphocytes isolated from peripheral blood samples using BD Vacutainer CTP cell collection tubes (BD). Total RNA was extracted using TRIzol (Invitrogen), purified with RNeasy Mini Kit (Qiagen), and hybridized to human Affymetrix U133 Plus 2.0 arrays (Affymetrix) according to manufacturer's instructions. RNA concentration was measured with a ND-1000 spectrophotometer (ThermoScientific) and the sample quality analyzed using the RNA Nano assay of the 2100 Bioanalyzer platform (Agilent Technologies) prior to cDNA synthesis. Two micrograms of total RNA was treated according to the eukaryotic RNA labeling protocol (Affymetrix). 15 micrograms of biotin labeled cRNA was fragmented according to the Affymetrix eukaryotic sample protocol. Hybridization, staining and washing of the chips was performed under standard conditions. The arrays were scanned with GeneChip Scanner 3000 7G (Affymetrix) at the Biomedicum Genomics core facility (Helsinki, Finland).

All gene expression data analysis and handling was performed using Bioconductor 2.3 (http://www.bioconductor.org) (Gentleman *et al.* 2004) implemented in R 2.8.0 software (http://www.r-project.org). Sequence-based re-annotation of the Affymetrix probes was performed according to the latest release of the Entrez gene database (build 36.3) (Dai *et al.* 2005). Re-annotation was performed, because it is widely known that the original selection of probes by Affymetrix relied on early, incomplete genome and transcriptome annotation. Thus, a probeset that is supposed to measure the expression level of a single gene might contain a significant

proportion of individual probes which map to multiple, or completely wrong transcripts, biasing the overall signal (Zhang *et al.* 2005). The original Affymetrix U133Plus2.0 chip set contains 54120 probe sets, whereas after re-annotation they were grouped in 17788 probesets, representing 17726 unique Entrez gene identifiers and 62 quality control probesets. The re-annotation packages for Bioconductor were obtained from the NuGO R-server (http://nugo-r.bioinformatics.nl/NuGO_R.html) and Brainarray server (http://brainarray.mbni.med.umich.edu/Brainarray/default.asp). Package versions 11.0.2 (database) and 11.0.1 (cdf and probe) were used. Extensive quality control (QC) of the raw data was carried out, in order to ensure a good array-array correlation. Basic QC was performed using the AffyQCReport package and the degree of RNA degradation was addressed using the AffyRNAdeg function of the Affy package implemented in Bioconductor.

Preprocessing of the data and calculation of the expression values was performed using the robust multiarray average (RMA) algorithm (Irizarry *et al.* 2003) implemented in the Affy package (Gautier *et al.* 2004) in Bioconductor. The RMA algorithm fits a linear model for all probesets across all arrays. Selection of the differentially expressed genes was performed with statistical methods implemented in the Limma package (Smyth 2004). Empirical Bayes method was used to moderate standard errors for the estimated log-transformed fold changes. This results in more stable inference and improved power, especially for experiments with small numbers of arrays (Smyth 2004). The basic statistic used to analyze the significance of differential expression is a moderated t-statistic, which is computed for each probeset and each contrast (in this study: only one contrast, cases versus controls). It is interpreted as ordinary t-statistic except that the standard errors have been moderated across genes using a Bayesian model. Also, the degrees of freedom are increased, reflecting the greater reliability associated with the smoothened standard errors (Smyth 2004). P-values were adjusted for multiple testing using the Benjamini and Hochberg's (BH) method (Benjamini and Hochberg 1995). Additionally, a B-statistic, i.e. the log-odds that a gene is differentially expressed, is produced. A B-statistic of zero corresponds to a 50-50 chance that the gene is differentially expressed, and it is automatically adjusted for multiple testing by assuming that 1% of the genes are expected to be differentially expressed.

The publicly available gene expression dataset used in Study III (referred to as AGRE-EXPR) was analyzed using TIGR 40K human arrays and obtained as normalized data which we further re-annotated and analyzed in the same way as the CF-EXPR data. Probe information was provided as Genbank IDs (total number of probesets=41472). However, in agreement with the original publication of the GSE15402 dataset (Hu *et al.* 2009b) we applied a 30% filter to the data, removing all probes with missing values in > 30% of the samples (n=35/116). We removed in total 16078 of 41472 probesets, leaving 25394 probesets for analysis. After removing probes with no matching Genbank ID, we were left with a probe list

corresponding to 15 920 Entrez gene IDs. Unlike in the original publication, where the individuals with autism were divided into three phenotypic subgroups based on ADI-R scores, we assigned all samples as cases or controls only (87 cases and 29 controls).

4.4.5 Pathway analysis

Pathway analysis was performed with a non-parametric in-house developed algorithm (named GWANA), which can be applied to both GWAS and gene expression datasets, as described before (Pietilainen *et al.* 2008, Aulchenko *et al.* 2009). The method uses the ranking of genes or transcripts in a dataset together with their biological pathway annotations to list pathways that are enriched among the most highly associated or differentially expressed genes. A cumulative score is calculated based on how extreme the observed combination of associated (i.e. high ranking) genes in a given pathway is. This score is compared to all associated or differentially expressed genes and their pathways. The more genes a pathway has among all top-ranked genes or transcripts, the better the cumulative score. The significance of this enrichment is inferred from the distribution of 10000 permutation cycles. All detailed description of the GWANA method is provided in the Supplementary Note of Study III.

The pathway analysis was based on the Gene Ontology (GO) classification of genes, with all three ontologies (biological process, molecular function, cellular component) included in the same analysis. The topology of the GO-tree was fully utilized by enumerating all available routes towards the root of the GO tree and adding all encountered vertexes as GO annotations for the given gene. The maximum size for a reported pathway was set to 200, an arbitrary cut-off aimed to limit the analysis to biologically more meaningful pathways. We required at least two genes to be associated or differentially expressed per pathway, in order to prevent small categories appearing to be significantly overrepresented on the basis of a single, possibly chance hit. Closest genes for SNPs and GO annotations for genes were queried from the Ensembl database (release 53, hg36).

GWAS datasets were analyzed by ranking all SNPs according to their original GWAS p-value. SNPs were mapped to genes by setting the maximum allowed distance of a SNP from its representative gene to 10kb from the 5’ and 3’ ends of a gene. When multiple SNPs mapped to a single gene, the SNP with the highest rank was retained, whereas SNPs that did not map to genes with the applied criteria were excluded. The permutation step was applied to the initial marker ranks, meaning that the SNP association p-values were randomized and the analysis repeated 10 000 times. In addition to evaluating the significance of the pathway enrichment, the permutation step also accounts for any gene length and uneven SNP distribution

related bias in the analysis. To find out how much the size of the input SNP list affects the results, we ran the pathway analysis with different p-value thresholds for individual SNPs ($p<0.05/0.01/0.005/0.001$). This was also done in order to avoid limiting the analysis on the most significant variants (or transcripts) only, since this approach would be likely to ignore numerous false negative hits, especially in a phenotype like autism, where effect sizes of common genetic variants are small.

Gene expression datasets were analyzed by ranking all transcripts based on i) p-value for differential expression and (ii) absolute fold change to allow for both up and downregulated genes in the same pathway. As with SNP data, initial transcript ranks were permuted 10 000 times, and if duplicate genes were present, the highest ranking of that gene was used. No p-value thresholds were applied, allowing the use of the whole input list in the analysis.

In order to further evaluate the performance of the GWANA pathway method we also applied it to GWAS and expression data from the AGRE study sample, as well as GWAS data from Crohn's disease (CD), for which the underlying biology is better established and certain pathways can be expected to be found.

5 RESULTS AND DISCUSSION

5.1 Role of *DISC1* in ASDs (Study I and unpublished data)

Since the initial report of the balanced translocation disrupting *DISC1* and co-segregating with schizophrenia and other major mental illnesses in a Scottish pedigree (St Clair *et al.* 1990), the gene has been extensively analyzed in many neuropsychiatric phenotypes. Following these, a wide range of functional studies has emerged (see Section 2.2.10), and *DISC1* has become one of the most studied candidate genes in psychiatric genetics. Due to (i) the established biological functions of the DISC1 protein in early neurodevelopment, (ii) the versatility of neuropsychiatric phenotypes associated or otherwise related to *DISC1*, and (iii) shared neurocognitive defects between schizophrenia and ASDs such as impaired executive function and social functioning (Baron-Cohen and Belmonte 2005, Happe *et al.* 2006), it seems possible that abnormal DISC1 functioning would underlie also early-onset neuropsychiatric disorders, such as autism and AS. This is supported by the sex-dependent association findings reported for *DISC1*, which is of interest regarding the overall higher prevalence of all ASDs in males. In this study, we set out to investigate the possible role of genetic variants in *DISC1* in autism spectrum disorders, with the hypothesis that DISC1 might be involved in fairly general neurodevelopmental processes, which, if disturbed, could lead to several slightly differing and even overlapping phenotypes such as ASDs and schizophrenia. Since one of the early association findings of *DISC1* and SCZ was reported in Finnish families (Hennah *et al.* 2003), we hypothesized that the probability of finding a common genetic variant for ASDs and SCZ might be higher in the isolated Finnish population compared to study samples with mixed backgrounds.

5.1.1 Association analysis

We tested in total 11 SNPs and two microsatellite markers on chromosome 1q42, spanning a ~600 kb region with *DISC1*, *DISC2*, and *TSNAX* genes. Family-based association of single markers was detected in the autism study sample only with D1S2709, a microsatellite marker intragenic to *DISC1* (TRANSMIT global p=0.022; FBAT p=0.010). This association was stronger when only affected males were included in the analysis (TRANSMIT global p=0.019; FBAT p=0.004). In the AS sample, no evidence of single marker association was seen. However, when only affected AS males were considered, SNP rs1322784 displayed association with both TRANSMIT (global p=0.0058) and FBAT (p=0.0195) (Table 9). This particular SNP is located ~101 kb apart from D1S2709, and appeared noteworthy also when comparing allele frequencies across study samples. The major allele frequency in the

AS sample was 0.83 compared to 0.78 in the regional controls and 0.72 in the Finnish population controls. In the combined study sample of both autism and AS families, modest association was seen with rs1411771 with FBAT (p=0.042; p=0.029 affected males only), but this could not be detected with TRANSMIT.

When analyzing Pedigree 1 separately, evidence of association was found for rs1322784, the best SNP in the AS families. For this SNP, the most significant association was observed when calculating LD under the assumption of linkage (LD | Linkage), with a p-value of 0.0007. The joint analysis of LD and linkage (LD + Linkage) yielded p=0.002. Modest evidence of association was detected also for D1S2709, the best marker in the autism families (p=0.0376, LD + Linkage; p=0.0260 LD | Linkage). Interestingly, when marker allele frequencies in Pedigree 1 ($n_{affected}$=33) were compared with other study samples, we noticed, that the major allele frequency in cases for the best SNP, rs1322784, was 0.98 compared with the frequency in the regional controls of 0.78 and the Finnish population average of 0.72. In fact, all except one of the affected individuals were AA homozygotes for rs1322784. This deviation between cases and both regional and Finnish population controls appeared highly significant using Fisher's exact test (p=9.3x10-5 and p=9.89x10-7, respectively).

Since linkage to 1q42 has been observed in Finnish SCZ families, we also analyzed our samples for two-point linkage. No significant evidence of linkage was observed in any of the study samples. However, again, some modest evidence emerged, when only AS males were considered (best LOD score 1.23, rs1000731, dominant model).

Table 9. ***Association analysis results of DISC1 in the autism and AS families.*** *P-values < 0.05 are marked in bold. The SNP alleles are given in column two, with the major allele listed first.*

Marker	Alleles	Autism families				Asperger syndrome families			
		FBAT all	Males only	TRANSMIT all	Males only	FBAT all	Males only	TRANSMIT all	Males only
rs1630250	G/C	0.823	0.444	0.649	0.289	0.138	0.220	0.147	0.303
rs1655285	G/C	0.847	0.549	0.456	0.272	1.000	0.739	0.912	0.506
D1S251		0.897	0.609	0.849	0.691	0.205	0.196	0.268	0.107
rs751229	T/C	0.553	0.972	0.796	0.737	0.889	0.402	0.766	0.282
rs3738401	G/A	0.426	0.197	0.788	0.522	0.423	0.170	0.324	0.237
rs1322784	A/G	0.279	0.542	0.250	0.408	0.086	**0.0195**	0.160	**0.0058**
rs967244	A/G	0.453	0.705	0.405	0.821	0.346	0.718	0.295	0.415
rs6675281	C/T	0.724	0.945	0.867	0.986	0.175	0.564	0.226	0.455
rs1000731	C/T	0.643	0.536	0.583	0.446	0.285	0.385	0.284	0.272
D1S2709		**0.010**	**0.004**	**0.022**	**0.019**	0.351	0.972	0.579	0.979
rs821616	A/T	0.783	0.592	0.568	0.271	0.701	0.403	0.469	**0.038**
rs1411771	T/C	0.106	0.050	0.076	0.077	0.299	0.290	0.471	0.591
rs980989	G/T	0.267	0.210	0.208	0.282	0.541	0.403	0.609	0.415

5.1.2 Haplotype association analysis

Haplotype association analysis was performed only to four specific haplotypes (named HEP1-4) which have previously shown association in Finnish families ascertained for schizophrenia. We tested the same haplotypes as in the original study, except in the case of HEP2 where rs1630250 and rs1655285 were used as surrogates for the haplotype information provided by the original haplotype (Figure 9).

In the autism sample and in the combined sample, none of the tested haplotypes showed evidence of association (best p-values 0.135 and 0.107, respectively). Suggestive association was detected only in the AS study sample with the HEP3 haplotype (rs751229 and rs3738401); no evidence of association was observed with HEP1, 2 or 4. In the whole AS sample, a global p-value of 0.030 was obtained for HEP3 with TRANSMIT (best allele combination TA; 17.6 observed transmissions, 12.7 expected transmissions). The same alleles were associated also when only affected AS males were used in the analysis (TRANSMIT global p=0.015; 13.5 observed transmissions, 9.0 expected transmissions). However, HBAT did not show any association with the exact same haplotype (Table 10).

In order to further define the extent of the potentially associating haplotype, we combined the HEP3 SNPs with neighboring SNPs in both directions. In the whole AS sample, none of the combinations (rs1655285 + HEP3 / HEP3 + rs1322784) appeared significant. However, again, when only AS males were considered, the combination of HEP3 plus the additional rs1322784 in the telomeric direction (HEP3+1) gave a global p-value of 0.0013 (TRANSMIT; best allele combination TAA; 13.4 observed transmissions, 9.0 expected transmissions). The HBAT analysis displayed a p=0.024 with two of these three SNPs, rs3738401 and rs1322784.

The most significant association signal originates from a ~150 kb region delimited by rs751229 and rs1322784. Similar association results at this region have been observed in many earlier studies, although not with the exact same HEP3+1 combination of SNPs. To name a few, Zhang and colleagues reported a haplotype association to schizophrenia involving the HEP3 SNPs and a third SNP preceding rs1322784 (Zhang *et al.* 2006), whereas another study reported an association of rs1322784 to schizophrenia and HEP3 to schizoaffective disorder (Hodgkinson *et al.* 2004). Two other studies reported associations of haplotypes involving various combinations of the HEP3 SNPs and rs1322784 with schizophrenia, bipolar disorder, and psychotic disorder (Thomson *et al.* 2005b, Palo *et al.* 2007). Overall, most association findings seem to localize to two distinct regions of DISC1, the 3' and 5'end of the gene, probably implying that the actual risk allele(s) remain to be identified or that multiple distinct genetic mechanism are operating on the region.

Since our sample size is relatively small, these results warrant replication in an independent ASD study sample. However, the fact that our finding co-localizes with the HEP3 region of the gene, which has been replicated several times to date, makes it highly interesting. Also, the associating alleles are the same that showed association in Finnish SCZ and BPD study samples. Especially interesting is the observation that in Pedigree 1, all but one of the affected individuals are homozygous for the best SNP in this study (rs1322784), possibly indicating enrichment of risk alleles in individuals with common genealogical origin.

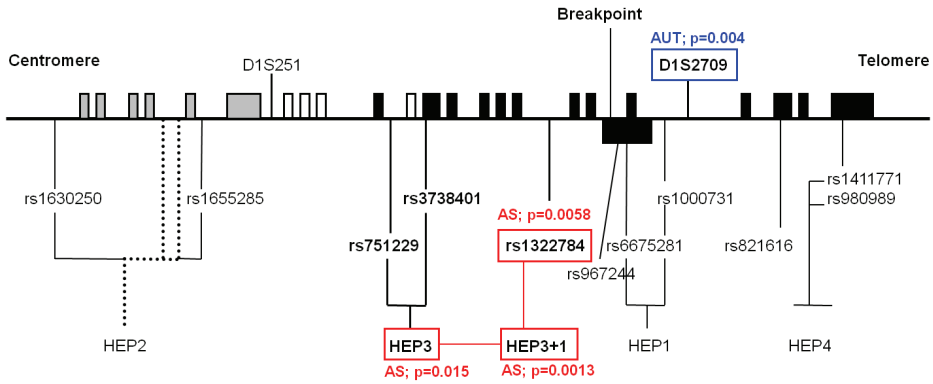


Figure 9. ***Schematic figure showing the exonic structure of the DISC1 region at 1q42.*** *DISC1 and DISC2 exons are marked in black, TSNAX exons in gray, and intergenic exons in white. Most significant single marker and haplotype associations in the autism study sample (blue) and AS study sample (red) are denoted. Non-synonymous SNPs are indicated with a star. The location of the original HEP2 haplotype (Hennah et al. 2003) is indicated with dotted lines.*

The association signal in this study seems to arise from a broad diagnostic category of ASDs, and from affected males in particular (although it should be noted that most of the affected individuals in all of the study samples are males, given the overall higher prevalence of ASDs in males). We have made an effort to carefully harmonize the diagnostic criteria across all families, and firmly establish the diagnosis, not only for autism but also for AS. To date, evidence of *DISC1* association and linkage has been observed in multiple neuropsychiatric phenotypes, in addition to schizophrenia (see Section 2.2.10). This is in line with the original *DISC1* translocation finding, which was not associated with schizophrenia only. The accumulating evidence strongly suggests a role for *DISC1* in a broad range of neurobiological and developmental processes, which are capable of causing

multiple, sometimes overlapping psychiatric conditions when disrupted. Interestingly, after the results of Study I were published, another study reported a deletion at 1q42 involving *DISC1*, *DISC2*, and *TSNAX* in an individual affected with developmental delay and autistic behaviours (Williams *et al.* 2009), providing further support for the involvement of *DISC1* in ASDs.

Table 10. ***Results of the DISC1 haplotype association analysis.*** *Haplotypes are presented in the genomic order. HEP3-1 and HEP3+1 were not tested in the autism study sample (x) because no evidence of association was seen in the initial analysis of HEP1-4. P-values < 0.05 are marked in bold.*

Haplotype	SNPs	Autism families				Asperger syndrome families			
		HBAT all	Males only	TRANSMIT all	Males only	HBAT all	Males only	TRANSMIT all	Males only
HEP2	rs1630250 rs1655285	0.912	0.519	0.492	0.242	0.244	0.463	0.173	0.143
HEP3-1	rs1655285 rs751229 rs3738401	x	x	x	x	0.578	0.212	0.308	0.143
HEP3	rs751229 rs3738401	0.707	0.592	0.954	0.704	0.288	0.115	**0.030**[a]	**0.015**[a]
HEP3+1	rs751229 rs3738401 rs1322784	x	x	x	x	0.197	0.670	0.110	**0.0013**[b]
HEP1	rs6675281 rs1000731	0.805	0.668	0.796	0.851	0.299	0.573	0.229	0.416
HEP4	rs1411771 rs980989	0.263	0.135	0.242	0.159	0.553	0.571	0.819	0.449

[a] Alleles TA
[b] Alleles TAA

5.1.3 Polymorphic miRNA target prediction

To search for biological explanations for the wide-ranged effects of the *DISC1* gene, we focused on SNPs located in a miRNA binding site, i.e. the seed region, since such SNPs can affect the binding and the regulatory relationship between a miRNA and its target gene in an allele-specific manner, and thus associate to the resulting phenotype. Abelson and colleagues (2005) were the first to show that a SNP in the 3' UTR of *SLITRK1* affected the interaction between the gene and human miRNA hsa-miR-189, and associated with Tourette's syndrome (see also Section 2.1.5). Likewise, Clop *et al* (2006) identified a G to A substitution in the 3' UTR of the sheep *GDF8* gene that created an illegitimate target site for two miRNAs, causing translational inhibition of the mutant transcripts.

SNP alleles can affect miRNA binding in two ways. They can either disrupt an existing target site leading to the loss of normal repression control of the target gene, or create a novel target site leading to abnormal target gene repression. The significance of the effect depends for example on the degree of conservation of the target site and redundancy effects. Since population genetic data supports strong purifying selection against SNPs that destroy conserved target sites and create novel, illegitimate targets (Hiard *et al.* 2010), the polymorphic miRNA target sites that do exist are likely to have true biological effects.

Since the available full length *DISC1* clone (IMAGE:900180) corresponds to the *DISC1* Long variant (Lv) isoform, we limited our search of polymorphic target sites to the two longest known *DISC1* isoforms, L and Lv. Both Patrocles and PolymiRTS identified three SNPs with potential effect on a predicted miRNA binding site (rs11122396, rs9308481, and rs11803088). However, the predicted miRNAs differed slightly between the two programs. In addition, PolymiRTS identified a fourth SNP, rs980989, with a possible effect on a miRNA binding site. We included all of these four SNPs and all of the nine targeting miRNAs to further functional analysis. In total, seven miRNAs were predicted to target the wild-type *DISC1* Lv construct. Two additional miRNAs were predicted to target the alternative SNP alleles (Table 11).

Of the four SNPs, rs980989 did not show association to ASDs in Study I (see Table 9). The other three SNPs were not included in the original study, but were subsequently genotyped in the same set of samples and tested for association. None of them showed evidence of association with either autism or AS ($p>0.05$, data not shown). However, there is association evidence for these SNPs in other neuropsychiatric phenotypes, such as schizophrenia and various cognitive traits (Hennah *et al.* 2003, Palo *et al.* 2007).

With two of these SNPs, rs11122396 and rs980989, the ancestral SNP allele, determined by human versus chimpanzee genome alignment, is not predicted to affect miRNA binding. Instead, in both cases, the derived allele is predicted to create a novel, illegitimate binding site and cause abnormal gene repression. With rs9308481, the ancestral allele G is part of a non-conservative binding site for four mature miRNAs, whereas the derived allele is predicted to disrupt this site. However, the SNP that is most likely to have a true effect is rs11803088, whose ancestral allele C is part of a conserved miRNA target site for hsa-miR-559. The derived allele T is predicted to disrupt this conserved site. If this were to be true, expression of *DISC1* in individuals TT homozygous for this SNP would be higher than in CC homozygotes, due to loss of normal repression control of hsa-miR-559. However, the situation is further complicated by the fact that the derived T allele is simultaneously predicted to disrupt a non-conservative binding site for hsa-miR-548c.

All predictions for miRNA binding and effect are blind to redundancy effects. A gene is typically targeted by multiple miRNAs, so the loss of the regulatory effects of just one might not have an effect at all, or it might be too small to be detected, given that the silencing effects of miRNAs are typically less than 50% (Bartel 2009). Also, the overall effect on target gene expression might be quite different in heterozygote versus homozygote individuals, who lack the binding site completely.

Table 11. ***Summary of DISC1 polymorphic miRNA target prediction results.*** *Predictions are combined from Patrocles and PolymiRTS. The SNP alleles of the wild type DISC1 Lv construct are indicated with an asterisk (*). miRNAs and SNPs predicted by both programs are underlined. The effect of the SNP on the target site is denoted as follows: C=derived allele creates a novel miRNA binding site, N=derived allele disrupts a non-conservative miRNA binding site, D=disrupts a conservative miRNA binding site.*

#rs	Position[a]	MAF CEU	Octamer ("seed")[b]	Effect	SNP allele	Targeting miRNA (stem loop)	Stem loop location	Mature miRNA
rs11122396	230241891	0.042	caAGCC[A]Taactc	C C C	A* G (ancestral)	miR-135a-1 miR-135a-2 miR-135b *none*	3p21 12q23.1 1q32	hsa-miR-135a hsa-miR-135b
rs980989	230242818	0.217	tttACA[T]TCAtat	C	T G* (ancestral)	miR-409 *none*	13q32.3	hsa-miR-409-3p
rs9308481	230242929	0.225	taaAAT[G]TGAagt aaaAT[G]TGAAgt AAT[G]TGAA	N N N N	G* (ancestral) A	miR-323 miR-23a miR-23b miR-130a *none*	14q32 19p13 9q22.3 11q12	hsa-miR-323-3p hsa-miR-23a hsa-miR-23b hsa-miR-130a*
rs11803088	230243392	0.069	cttTTA[C]TTTtaa cttttA[T]TTTTAa	N D	C* (ancestral) T	miR-559 miR-548c	2p21 12q14.2	hsa-miR-559 hsa-miR-548c-3p

ABBREVIATIONS: MAF=minor allele frequency, CEU=HapMap CEPH population

[a] RefSeq build 36.3

[b] Genomic sequence 5' to 3'

5.1.4 Effect of miRNAs on *DISC1* expression

We tested the effects of altogether seven miRNAs on endogenous *DISC1* expression in 293FT cells and on the overexpression of *DISC1 Lv*, hsa-miR-135a, 135b, 323-3p, 23a, 23b, 130a*, and 559. We started with these seven because they target the wild type alleles of rs11122396, rs980989, rs9308481, and rs11803088 of the *DISC1 Lv* construct (Table 11). In qPCR, for all samples, the standard deviation of the Ct values of the technical replicates was < 0.5, with most samples having a standard deviation of < 0.3, indicating good technical quality. Variation among biological replicates was more substantial, especially in the overexpression experiments.

In the endogenous experiment, based on the linear regression analysis, two miRNAs (hsa-miR-135b and hsa-miR-559) significantly reduced the level of endogenous *DISC1* expression compared with the negative control miRNA (p-values 0.015 and p=0.0045, respectively). The two siRNAs (positive controls) also displayed significant downregulation of endogenous *DISC1* (p-values 0.016 and 0.023), as expected (Table 12).

In the overexpression experiment, significantly more variation and "noise" was seen among the biological replicates. Based on the regression analysis, no miRNA induced a statistically significant effect on *DISC1* expression, and of the positive controls, only one produced a significant reduction in *DISC1* expression (p=0.032). However, miR-559 and miR-135b still showed the same direction of effect as in the endogenous experiment in most of the biological replicates (Table 12). Since the overexpression assay allows the testing for possible allele specific effects, we repeated the experiment for these two miRNAs with the mutated *DISC1 Lv* constructs harboring the alternative alleles for the two SNPs these miRNAs target. Very interestingly, as predicted, with miR-559 the expression level of *DISC1* increased notably when using a construct with the derived SNP allele for rs11803088, compared with the ancestral allele. This would suggest that the normal repression control mediated by miR-559 is lost with the derived allele. With miR-135b, the allele-specific effects were not as clear, and additional replicates are required to clarify the effect.

It should be noted, that endogenous *DISC1* expression includes the total expression of all *DISC1* isoforms targeted by the primer, whereas the overexpression experiments primarily measure effects on the Lv isoform, since the overexpression from the used construct is driven by a strong promoter compared with endogenous *DISC1* levels. Since the multitude of *DISC1* transcript variants remain poorly characterized, it is possible that the introduced miRNAs have effects also on the other isoforms. Most of the *DISC1* isoforms differ with regard to their 3'UTR sequence, meaning that they are likely to be regulated by different sets of miRNAs.

However, it is not possible to address isoform specificity (other than Lv) in this study setting.

There are numerous factors affecting our experiment which need to be tested before any final conclusions can be made. For example, to verify the robustness of the silencing effect, the experiments need to be repeated with different numbers of cells and different miRNA concentrations, and the level of endogenous miRNA expression in 293FT cells needs to be determined for the studied miRNAs. Also, it should be investigated, whether the observed effects can be reversed by knocking down the miRNA mediating the effect. Even if a significant reduction in *DISC1* expression is seen when a specific miRNA is introduced, it is hard to prove that the effect is in fact mediated by the miRNA silencing pathway of the cell. It is possible that the miRNA actually binds somewhere else, and the silencing effect is caused as a secondary, downstream effect. Thus, knocking down the endogenous miRNA, and possibly also the miRNA overexpression, to reverse the effect would provide an additional level of confidence. Further, since it is well-recognized that miRNA silencing effects are not always consistent on the RNA and protein levels, the effects on *DISC1* expression should be measure also on protein level. Thus, the next step in our study will be to perform the same set of transfection experiments using luciferase assays, which is a quick, antibody-independent way of monitoring changes at the protein level.

Table 12. ***Summary of the qPCR results.*** *Results are presented for the two best miRNAs and two positive controls in three independent experiments. Additionally, the results for the alternative allele in the overexpression experiment are presented in two independent experiments. Fold change values and the percent difference are calculated relative to the negative control miRNA. The reported values represent the mean of three biological replicates.*

miRNA	**Experiment 1**		**Experiment 2**		**Experiment 3**		**Experiment 4**		**Experiment 5**	
	FC	**%diff**	**FC**	**%diff**	**FC**	**%diff**	**FC**	**%diff**	**FC**	**%diff**
Endogenous experiment										
miR-135b	-1.245	-24.5	-1.217	-21.7	-1.044	-4.4	NA	NA	NA	NA
miR-559	-1.204	-20.4	-1.330	-33.0	-1.199	-19.9	NA	NA	NA	NA
siRNA-A	-1.190	-19.0	-1.227	-22.7	-1.229	-22.9	NA	NA	NA	NA
siRNA-B	-1.251	-25.1	-1.295	-29.5	-1.197	-19.7	NA	NA	NA	NA
Overexpression experiment, wild type allele							*Alternative allele*			
miR-135b	-1.045	-4.5	-1.195	-19.5	-1.403	-40.3	1.091	9.1	-1.253	-25.3
miR-559	-1.255	-25.5	1.042	4.2	-1.463	-46.3	1.141	14.1	1.527	52.7
siRNA-A	-1.453	-45.3	-1.255	-25.5	-1.687	-68.7	-1.407	-40.7	-1.228	-22.8
siRNA-B	-1.214	-21.4	-1.350	-35.0	-1.183	-18.3	x	x	x	x

ABBREVIATIONS: FC=fold change, %diff=percent difference, NA=not applicable, X=not tested

5.1.5 Conclusions

Identifying the first reported association between *DISC1* and ASDs in a widely replicated genomic region is highly interesting, and clearly demonstrates that phenotypic borders in psychiatric genetics are loose. Considering the well-recognized challenges in diagnostics of mental disorders, it is likely that there are genes such as *DISC1*, which are involved in widespread neurobiological processes, which can be disturbed by various mechanisms leading to slightly differing, and even overlapping phenotypes. The existence of a common "*DISC1* pathway", involving *DISC1*-interacting genes, has already been suggested, providing a possible biological link between such phenotypes. *DISC1* can also be connected to ASDs via one of its interacting proteins, microtubule-associated protein 1A (MAP1A) (Morris *et al.* 2003). MAP1A physically interacts with discs, large homolog 4 (Drosophila) (DLG4) (Brenman *et al.* 1998) that is known to interact with neuroligin genes (Irie *et al.* 1997). Neuroligins are neuronal cell-adhesion molecules and well-known susceptibility genes for ASDs, proving further evidence of the molecular interconnectivity in neuropsychiatric disorders.

In this study, we have followed up our association finding and started to explore the possible molecular mechanisms through which *DISC1* might exert its widespread effects. We identified polymorphic miRNA target sites in the 3'UTR of *DISC1* and identified a specific miRNA, hsa-miR-559, which seems to have an allele-specific effect on *DISC1* expression *in vitro*. The study is still ongoing and further experiments are being conducted to obtain more information about the possible allele-specific regulatory relationship between *DISC1*, hsa-miR-559, and hsa-miR-135b. Although it is not possible to make final conclusions yet, altered miRNA regulation would offer a tempting explanation for some of the wide-ranged effects of *DISC1*.

5.2 Genome-wide linkage and LD in Pedigree 1 (Study II)

In this study, we described an extended ASD pedigree originating from Central Finland (referred to as Pedigree 1) (Figure 5), and analyzed it for ASD susceptibility loci. The pedigree consists of 20 families, scattered all over Finland, which have been genealogically traced back to the 17th century and found to originate from common ancestors. Since the genetic heterogeneity of ASDs has greatly hampered the identification of genetic risk loci and variants, we wanted to take advantage of the substantially higher degree of LD observed in population isolates such as Finland (Service *et al.* 2006) (see Section 2.1.4), and of the reduced genetic heterogeneity in genealogically connected individuals. Thus, we performed a

traditional genome-wide scan for linkage and LD in Pedigree 1 with microsatellites. We hypothesized that if the observed genealogical links of Pedigree 1 reflect identical-by-descent (IBD) sharing, a few causative variants (or even a single variant) could be expected to be enriched to this pedigree.

The pedigree was discovered when the genealogy of all of the families in the nationwide Finnish ASD study sample was systematically examined. First, ten families whose ancestors originated from a single small farm in Central Finland ~5-10 generations ago were identified. When the ancestral trees were followed back up to 12 generations, we were able to distinguish nine ancestors connecting these 10 families, which, most interestingly, were born on the same small farm 215-350 years ago (Auranen *et al.* 2003). Ten additional nuclear families with ASDs were later linked to this core pedigree. We were also able to reveal further genealogical links among the 20 families.

5.2.1 Linkage and LD scan

In the initial genome-wide scan, we analyzed 1109 microsatellite markers across the genome. Altogether nine loci were identified in the recessive LD+Linkage analysis of Pseudomarker (Table 13). These nine loci exceeded the threshold of –log(p)=2.5 which was chosen as the cut-off for the selection of a reasonable number of follow-up loci based on the overall distribution of the results (Figure 2 of Study II). To demonstrate that some of the signals were obtained from haplotype sharing among affected individuals instead of linkage only, we also monitored the LD|Linkage test of Pseudomarker for these nine loci, which allows for linkage but does not assume it. The most significant LD+Linkage results were obtained with D1S2707 on chromosome 1q23.2 and D15S156 on 15q12 (p=0.00082 and p=0.00081, Set1, respectively). With both markers, the evidence was almost entirely attributable to sharing across families (LD|Linkage, p=0.00079 and p=0.0009, respectively). In the dominant Pseudomarker analysis, only one locus exceeded the chosen threshold.

Additionally, one significant locus was identified in the multipoint linkage analysis with Simwalk2 (NPL_dominant, –log(p)=3.57, D19S591) at chromosome 19p13.3 (Figure 10). The second most significant locus in the Simwalk2 analysis was observed at chromosome 6 (NPL_dominant, D6S958, -log(p)=2.15, whilst results for all other loci were below –log(p)=1.5 with both dominant and recessive statistics (data not shown).

Table 13. ***Results of the initial genome-wide Pseudomarker analysis in Pedigree 1.*** *All markers exceeding –log(p)=2.5 in a Set1 analysis are listed.*

Ranking	Marker	Chr[a]	Genetic location[b] (cM)	LD+Linkage [-log(p)]	LD|Linkage [-log(p)]
Recessive Pseudomarker					
1	D15S156	15q12	15.1	3.09	3.05
2	D1S2707	1q23	156.1	3.09	3.10
3	D13S232	13q12	8.7	3.01	2.85
4	D14S283	14q11	14.7	2.96	2.52
5	D8S1132	8q23	~113.1	2.68	2.84
6	D6S1279	6p24	~30	2.66	1.59
7	D5S2090	5q32	150.0	2.62	2.81
8	D5S2006	5q35	205.7	2.56	2.74
9	D6S422	6p22	42.8	2.51	2.58
Dominant Pseudomarker					
1	D14S1071	14q12	28.2	2.76	2.93

[a] UCSC Human Genome Browser, May 2004 assembly (hg17)
[b] From ptel based on deCODE Genetic map (Kong *et al.* 2002)

5.2.2 Follow-up and candidate gene analysis

We chose altogether ten loci from the initial genome-wide scan for follow-up, including the nine most significant loci from the recessive Pseudomarker analysis and the one significant locus from the multipoint linkage analysis. Additional 44 microsatellites were genotyped and analyzed at these ten loci (full marker information and results are presented in Supplementary Table S1 and S2 of Study II). Based on the information from the follow-up results and previous evidence in ASDs for two of the loci, we chose three loci to take further to finemapping (1q23, 15q13, and 19p13). Chromosome 1q23 has been implicated by two previous genome-wide screens for ASDs in Finnish families (Auranen *et al.* 2002, Ylisaukko-oja *et al.* 2004), whereas prior evidence for 15q11-q13 arises from cytogenetic studies (Gillberg 1998, Wassink *et al.* 2001a, Veenstra-VanderWeele and Cook 2004). Regional candidate genes at these loci were chosen based on biological relevance to ASDs and previous research results, and analyzed using a total of 152 SNP markers (Table 14). Due to prior evidence in ASDs, SNPs on chromosome 1 and 15 were analyzed also in the nationwide ASD family sample, whereas SNPs on chromosome 19 were primarily analyzed in Pedigree 1, except for the most significant SNPs. The results of the candidate gene analysis are presented in full in Supplementary Table S2 of Study II.

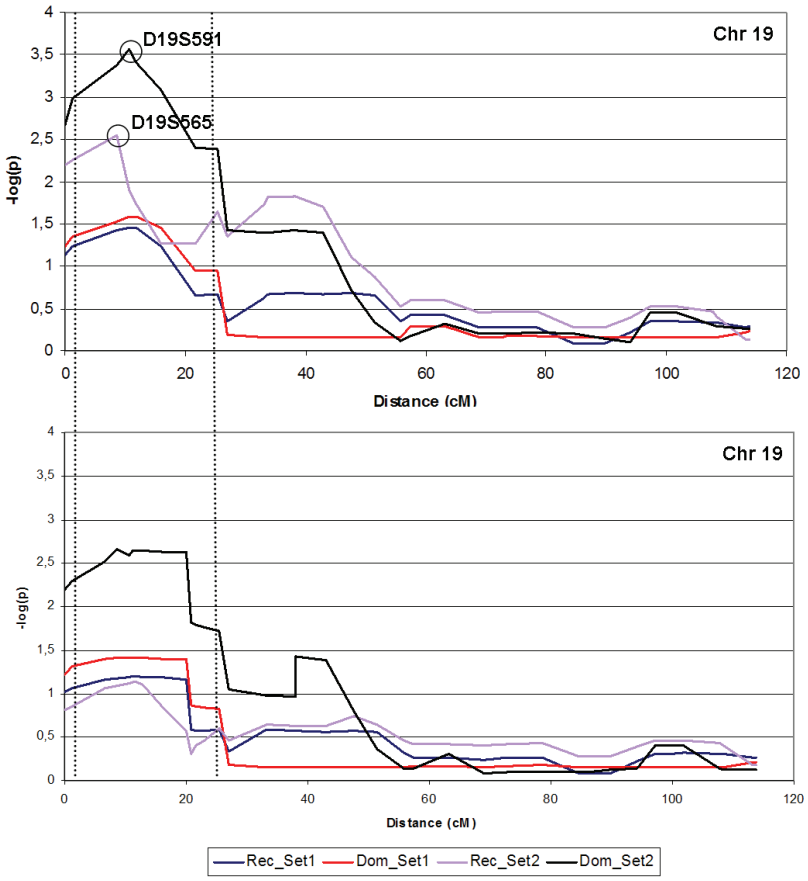


Figure 10. ***Multipoint linkage analysis results on chromosome 19 in Pedigree 1.*** *The upper figure shows the results of the initial genome-wide scan with the two best markers displayed, whereas the lower figure presents the signal after the inclusion of seven follow-up microsatellites. The follow-up region is indicated with dotted lines. Set1=analysis with only one affected individual from Family 2, Set2=analysis with all 13 affected individuals from Family 2. Results produced by Simwalk2. Abbreviations: Rec=recessive, Dom=dominant.*

On chromosome 1q, the finemapping covered 31 SNPs from six genes over a region of ~3 Mb. Most significant evidence was detected in Pedigree 1 at *ATP1A2* gene with rs1016732 (p=0.00048, Set2, LD+Linkage, recessive Pseudomarker), which is located just 14.7 kb away from the best microsatellite of the initial scan. As with the microsatellite, the association evidence was primarily attributable to sharing across the families (p=0.00058, LD|Linkage). Encouragingly, also the four consecutive SNPs were modestly associated (from p=0.03 to p=0.006, LD+Linkage). However, only marginal evidence of association was observed with any of these SNPs in the

nationwide collection of ASD families (best p-values ~0.01). *ATP1A2* is part of a syntenic rodent epilepsy locus, centered around D1S2707, together with *ATP1A4*, *KCNJ10*, and *KCNJ9* (Buono *et al.* 2004, Ferraro *et al.* 2004). Association between seizure susceptibility, idiopathic generalized epilepsy and *KCNJ10* has also been detected in humans (Lenzen *et al.* 2005), which is of interest, since up to 30% of individuals with autism suffer from epilepsy (Gillberg and Billstedt 2000).

On 15q, we analyzed 41 SNPs from four genes over a region of ~2.3 Mb. Modest association was detected with six SNPs from a $GABA_A$ receptor subunit gene cluster (from p=0.02 to p=0.0023, LD+Linkage, Set2, recessive Pseudomarker), with the most significant results originating from four consecutive SNPs within *GABRB3* (from p=0.04 to p=0.00084, LD|Linkage, best SNP rs7173713). As with the SNPs on chromosome 1q, only marginal evidence was seen outside Pedigree 1 in the nationwide samples.

Candidate genes on chromosome 19p were mainly analyzed in Pedigree 1 in the absence of previous evidence for this locus. Since the most significant multipoint linkage signal for this locus was obtained in a Set 2 analysis (Figure 10), with all 13 affected individuals from Family 2 included, we focused on the Set 1 analysis in the finemap, to identify association signals independent of Family 2 (see Section 4.1.3). We analyzed 80 SNPs from 13 genes over a region of almost 6 Mb. The most significant association in the whole study was observed within a cluster of three genes, *TLE2*, *TLE6*, and *AES* (also known as *TLE5*), located just 12.8 kb away from the best microsatellite at this locus, D19S591. Altogether eight consecutive SNPs yielded LD+Linkage p-values < 0.04 in the same analysis (Set1, dominant Pseudomarker), consistently with the original linkage. Of these SNPs, best results were obtained with rs4806893 and rs216283 (both p=0.000078), and rs216276 (p=0.00063). Evidence of sharing across families was observed with five of these SNPs (from 0.05 to p=0.00019, LD|Linkage). The eight SNPs cover a region of 16.5 kb and are located within *TLE2* and the 3'UTR/intergenic region of *TLE6* (the genes are transcribed in reversed directions). To further assess the role of these SNPs, we analyzed them further in the nationwide ASD family sample, but again, no comparable evidence of association was seen outside Pedigree 1. Based on haplotype analysis of the eight most significant SNPs, no single haplotype could be expected to account for the entire association signal (Supplementary Table S3 of Study II).

The three genes belong to the human TLE (transducin-like enhancer of split) family of proteins homologous to the *Drosophila* Groucho protein, which is involved in neurogenesis during embryonic development as part of the Notch signaling pathway (Stifani *et al.* 1992, Miyasaka *et al.* 1993). All of the human TLE-genes share a conserved TLE_N protein domain (Pfam ID PF03920) and act as transcriptional co-repressors. Similar to their *Drosophila* counterparts, human TLEs are thought to

negatively regulate neuronal development and differentiation (Chen and Courey 2000). Interestingly, loss-of-function of Groucho and other components of the Notch pathway result in the overproduction of neurons (Heitzler *et al.* 1996), which is of relevance given that macrocephaly and increased brain volume are frequent observations in individuals with autism (Fombonne *et al.* 1999, Cody *et al.* 2002). Interestingly, *TLE2*, together with *forkhead box G1* (*FoxG1*), was recently shown to be crucial in the formation of the ventral telencephalon (Roth *et al.* 2010).

5.2.3 Conclusions

Focusing on rare forms of common diseases has proved to be of high importance in unraveling genes and underlying disease mechanisms (for e.g. Vionnet *et al.* 1992). Also, the syndromic forms of autism spectrum disorders such as autism caused by the Fragile X mutation or a chromosome 15q11-q13 duplication, have received considerable attention, and increased our understanding of the underlying genetic and biological mechanisms. In addition, the autism field has recently seen many studies, where a single, high-penetrance mutation has been identified as the causative mutation in individual families (see Section 2.2.9). Since LD-based mapping has been successfully used to identify disease genes in monogenic diseases in the Finnish population (for e.g. Varilo *et al.* 1996), this study had an ideal setting for genetic mapping based on linkage and haplotype sharing, and the enrichment of a few causative variants in the affected individuals of Pedigree 1 was considered likely.

Unexpectedly, this was not what we found. The results of the genome-wide scan and the following candidate gene and haplotype analysis clearly showed that there are multiple genomic loci affecting the phenotype in the pedigree. Our results provide additional support to two previously established ASD risk loci, at 1q23 and 15q11-13, and highlight a third interesting locus at 19p13. Suggestive linkage at 19p13 has been reported in previous genome-wide linkage scans in ASDs, but the reported LOD-scores have been small. Thus, the multipoint linkage signal obtained in this study with just 20 families is promising. In fact, haplotype analysis at the linked regions revealed that at this locus, all of the families of Pedigree 1 show complete segregation with the trait (i.e. linkage), implying that this locus is likely to contribute to the phenotype in the pedigree. However, the locus appears specific to Pedigree 1, since none of the candidate gene association results replicate outside of the pedigree in the nationwide ASD study sample. Also, none of the candidate gene association signals were significant enough to completely explain the linkage signal, suggesting that additional variation is contributing to the signal at each locus. Given that this study was mostly performed with microsatellite resolution with an average intermarker distance of 3.43 cM, it cannot be ruled out that the affected individuals would share shorter regions of their genome in common. Study III was initiated to

address this question and follow-up the results of this study. However, if a shared rare variant would be present in the pedigree, it would have been found already by linkage.

Table 14. ***Candidate genes on 1q23, 15q12, and 19p13 chosen for finemapping.***

Gene	Gene name[a]	Chr	Size (kb)	Genotyped SNPs
KCNJ10	potassium inwardly-rectifying channel, subfamily J, member 10	1q23	32.0	6
KCNJ9	potassium inwardly-rectifying channel, subfamily J, member 9	1q23	7.9	2
ATP1A2	ATPase, Na+/K+ transporting, alpha 2 polypeptide	1q23	27.8	5
ATP1A4	ATPase, Na+/K+ transporting, alpha 4 polypeptide	1q23	35.4	7
NOS1AP	nitric oxide synthase 1 (neuronal) adaptor protein	1q23	298.6	6
RGS4	regulator of G-protein signaling 4	1q23	7.2	5
UBE3A	ubiquitin protein ligase E3A	15q12	68.3	10
GABRB3	gamma-aminobutyric acid (GABA) A receptor, beta 3	15q12	227.5	19
GABRA5	gamma-aminobutyric acid (GABA) A receptor, alpha 5	15q12	113.6	6
GABRG3	gamma-aminobutyric acid (GABA) A receptor, gamma 3	15q12	652.5	6
PALM	paralemmin	19p13	39.3	9
GRIN3B	glutamate receptor, ionotropic, N-methyl-D-aspartate 3B	19p13	9.3	2
EFNA2	ephrin-A2	19p13	13.8	2
MBD3	methyl-CpG binding domain protein 3	19p13	16.0	4
GNG7	guanine nucleotide binding protein (G protein), gamma 7	19p13	191.4	21
TLE6	transducin-like enhancer of split 6 (E(sp1) homolog, Drosophila)	19p13	17.6	7
TLE2	transducin-like enhancer of split 2 (E(sp1) homolog, Drosophila)	19p13	31.4	11
AES	amino-terminal enhancer of split	19p13	9.5	8[b]
GNA15	guanine nucleotide binding protein (G protein), alpha 15 (Gq class)	19p13	27.5	8
SH3GL1	SH3-domain GRB2-like 1	19p13	40.1	4
SEMA6B	sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6B	19p13	15.9	2
NRTN	neurturin	19p13	4.5	1
PSPN	persephin	19p13	0.6	1

ABBREVIATIONS: Chr=chromosome, bp=base pair

[a] According to HUGO Gene Nomenclature Committee (HGNC)

[b] Includes one intergenic SNP

5.3 The genetic architecture of ASDs in genealogically connected individuals (Study III)

In Study III, we used three complementary approaches to follow up the results obtained in Study II. First, we extended our study sample (Pedigree 1, 18 families) with 33 additional families originating from the same geographical region (including Pedigree 2) (referred to as CF-GWAS). Second, we increased our marker density from microsatellites to genome-wide high-density SNP data. This enabled us to investigate whether the affected individuals in Pedigree 1 and Pedigree 2 would share shorter genomic regions than visible with microsatellite resolution. Third, we performed pathway analysis using information from both GWAS and global gene expression analysis from the same study sample. With these approaches we aimed to thoroughly dissect the genetic background of ASDs in this population sub-isolate, and begin to explore the biological processes and mechanisms underlying ASDs in these individuals.

5.3.1 Genome-wide association analysis

Isolated populations have traditionally been useful in mapping of monogenic disorders, and recently, their usefulness also in complex disease genetics has been demonstrated. For example, a study in multiple sclerosis showed that if there are common risk factors affecting the phenotype enriched in individuals from an isolate, these can be identified using a very small number of distantly related individuals and carefully matched population controls (Jakkula *et al.* 2010). This study setting is analogous to the one here, which initially motivated us to perform a traditional case-control GWA analysis in our study sample, which in more admixed populations would be much too small to have enough statistical power to reliably identify common variation.

In the GWA analysis of the CF-GWAS study sample and matched controls (51 cases, 181 controls), we detected seven SNPs from seven different loci, with p-values $< 1 \times 10^{-5}$. One of the SNPs, rs9309326 on chromosome 2p16.1, reached genome-wide significance with $p=6.88 \times 10^{-9}$. None of these SNPs overlapped with previously reported autism GWAS results, even though two were located on chromosome 5p, a region implicated by two previous studies (Wang *et al.* 2009b, Weiss *et al.* 2009). None of the identified loci overlapped with the linkage regions from Study II either, but this is not unexpected since a GWA analysis is specifically designed to target common genetic variants, unlike linkage analysis. To rule out possible sources of bias regarding this SNP, we took a careful look at the quality control measures. However, no genomic inflation was observed in the genome-wide distribution of p-values ($\lambda=1.05$), which rules out possible population stratification effects. The success rate for the SNP was 99.6%, HWE p-value p=0.0001, and the

three genotype categories were clearly separated into clusters, indicating good overall genotyping quality for this SNP. The odds ratios (OR) of rs9309326 and the six other significant SNPs are very large compared with previously reported results from autism (ORs from 3.13 to 7.08), but it is not unexpected to observe larger effects in isolated populations. Alternatively, the effect sizes might be overestimated due to small sample size.

The minor allele frequency (MAF) for the most significant SNP was 0.28 in the CF cases versus 0.07 in the controls. To assess the validity of the allele frequency in the relatively small set of controls used, we checked the MAF also in a set of 7740 samples from the general Finnish population. The MAF was 0.11, compared with 0.14 in the HapMap CEU population, suggesting good comparability. The association signal was not traceable back to a single pedigree, as the minor allele was carried by 50% of the cases in Pedigree 1 (n=9), one case in Pedigree 2, and 14 of the other CF cases. To further assess the role of the most significant GWAS findings, all SNPs with $p<1 \times 10^{-4}$ in the CF-GWAS (36 in total) were genotyped in all available family members of the CF-GWAS cases, as well as in the nationwide collection of Finnish autism families (n=126). A family-based analysis in these individuals did not replicate the signals from the original case-control results, suggesting that they are specific to the CF isolate.

5.3.2 Haplotype analysis

The genome-wide significant SNP rs9309326 is located between two recombination hotspots, based on recombination rates estimated from HapMap data (Figure 1 of Study III). To further characterize the region surrounding this SNP, we performed haplotype association analysis with ten-SNP sliding windows at this locus. As expected, the most significant association was seen with a haplotype which includes rs9309326 ($p=9.96 \times 10^{-7}$), but we also observed three other haplotypes with significant associations ($p<3.90 \times 10^{-5}$), suggesting that the association signal at this locus is not driven entirely by rs9309326. Also, the association signal started to decrease as soon as SNPs exceeding the recombination hotspots were included in the haplotypes.

The lack of association with other SNPs at the region can in part be explained by the low correlation (i.e. degree of LD) among the SNPs ($0.01 < r^2 < 0.13$), as expected with 317k SNP chip data. SNPs on this platform have specifically been chosen to tag single haplotype blocks only, thereby being mostly uncorrelated. However, modest LD could still be seen with rs9309326 and the haplotypes not including the SNP. For all haplotypes, the global association signal could be attributed to haplotypes enriched to cases (frequency from 0.08 to 0.22) and almost absent from controls (frequency from 0.003 to 0.05).

The closest gene to the most significant SNP is *B-cell CLL/lymphoma 11A isoform* (*BCL11A*), located ~53 kb upstream. All but two of the associated haplotypes overlap with the 3' end of the gene, suggesting that both the SNP and the haplotypes tag an ASD risk allele at this locus. *BCL11A* has been strongly associated with regulation of fetal haemoglobin levels (Menzel *et al.* 2007, Sankaran *et al.* 2008, Uda *et al.* 2008). It is a TF specifically expressed in hematopoietic tissue and the brain (Leid *et al.* 2004). Interestingly, BCL11A was recently shown to alter the distribution of nuclear actin and downregulate axon branching in hippocampal neurons (Kuo *et al.* 2009). It has also been reported to interact with *Calcium/calmodulin-dependent serine kinase* (*CASK*), a causative gene for X-linked mental retardation and brain malformalities (Najm *et al.* 2008, Tarpey *et al.* 2009). CASK was shown to regulate axonogenesis through interaction with BCL11A (Kuo *et al.* 2010).

5.3.3 Shared segment analysis and homozygosity mapping

Genome-wide homozygosity and shared segment analyses were performed in Pedigree 1 and Pedigree 2, both of which have genealogical roots in Central Finland. Homozygosity mapping was used to test for possible recessive susceptibility variants whereas the shared segment analysis can identify shared, enriched risk variants inherited in a dominant-like fashion. The loci identified in these two analyses are summarized in Table 15. No regions shared by all of the affected individuals were identified in either pedigree, an observation made already with microsatellites in Study II with Pedigree 1. However, this study confirmed that even with a substantially denser resolution, no single genomic region is shared by all of the cases, despite distant relatedness. Instead, we identified small subsets of cases which shared regions in common. The most interesting regions of homozygosity (ROH) were identified on chromosomes 4p15.1 and 18q22 in Pedigree 1 and Pedigree 2, respectively. At these loci, all but one of the affected individuals per pedigree were homozygous for a small region (< 100 kb). The ROH on 4p15.1 does not overlap with any genes, whereas the region on 18q22 overlaps with a single gene *coiled-coil domain containing 102B (CCDC102B)* with a largely unknown function.

Similar results were obtained in the shared segment analysis. No regions were shared by more than four pairs of affected individuals (Study III, Supplementary Note). Altogether five regions of interest were identified, three in Pedigree 1 and two in Pedigree 2, but in all of these regions, there were typically only two pairs of cases (and no more than three), which shared the exact same allelic combination IBD. To find out whether the identified shared segments and ROHs were specific to the two pedigrees, we assessed the regions also in the matched controls. All of the regions were common in controls as well, suggesting that they represent ancestral

haplotypes and any possible mutations would have been introduced more recently onto this background. These results support the idea that risk factors for ASDs are family specific, although in theory, the small ROHs shared by almost all affected individuals in both pedigrees might still reveal a shared risk variant if sequenced.

Table 15. ***Regions of interest identified in the homozygosity and shared segment analyses in Pedigree 1 and Pedigree 2.***

Chromosome	PED	Start (kb)[a]	End (kb)[a]	Size (kb)	SNPs	Type
2q14.3	PED1	124193	129108	4915	74	Shared seg
4p15.1	PED1	33720	34094	373	23	ROH
4q34.1	PED1	173091	175514	2423	53	Shared seg
6p22	PED2	28629	28738	109	6	ROH
7q31.31	PED2	117503	117943	440	48	ROH
8q11.22	PED2	51513	52109	596	38	ROH
9q12-q21	PED2	68191	72418	4227	68	Shared seg
10p15.3	PED1	1433	2080	646	34	Shared seg
11p11.2	PED1	48105	48573	467	29	ROH
15q23-q24.1	PED1	69899	70794	894	57	ROH
15q23-q24.1	PED2	69889	70501	611	47	ROH
16q22.3	PED2	71027	71875	848	21	Shared seg
18q22.1-q22.2	PED2	64802	64929	126	37	ROH

ABBREVIATIONS: PED=pedigree, kb=kilobase, ROH=region of homozygosity, seg=segment
[a] According to the hg18 genome build.

In summary, using genome-wide SNP markers, we did not identify any shared genomic regions in the two CF pedigrees. Thus, it seems that even within a set of distantly related families, genetic heterogeneity is significant, and rare genetic variants specific to subsets of, or individual, families comprise the majority of risk factors for ASDs. None of the identified regions overlap with the linkage regions identified in Study II, which suggests that the putative rare variants harboured by the linkage regions in Pedigree 1 are distinct from the ones highlighted by the SNP-based shared segements and regions of homozygosity. Since many of the ROHs and shared segments contain genes that have been previously implicated in ASDs, such as *CNTNAP5* (Pagnamenta *et al.* 2010) and ankyrin-repeat domain genes (Marshall *et al.* 2008, Willemsen *et al.* 2010), the next obvious step is to sequence these regions to identify the underlying mutations. We additionally identified one genome-wide significant common SNP in the GWA analysis of the CF-GWAS dataset, which seems to confer risk to ASDs in these individuals.

5.3.4 Global gene expression analysis

Gene expression studies in autism have been greatly hindered by the availability of samples, and only a handful of studies have been reported. Yet, despite being small, these studies have established that expression profiling, most commonly from lymphoblastoid cell lines (LCLs), can distinguish between affected individuals and their healthy siblings, as well as between different syndromic forms of autism (Baron *et al.* 2006a, Nishimura *et al.* 2007, Hu *et al.* 2009b), thus again highlighting the importance of accurate phenotypic subgrouping in genetic studies.

Using mononuclear lymphocytes, altogether 325 differentially expressed genes were observed between ten ASD cases and ten controls (CF-EXPR dataset) at p=0.01 significance level (non-adjusted t-test p-value < 0.01). Of these, we defined 17 genes upregulated (FC > +1.5) and 55 downregulated (FC < -1.5) in individuals with ASDs compared to controls. The fold change range of these genes was -2.24 to +4.14. Ten most significantly up and downregulated genes are listed in Table 16. At p=0.05 significance level, 1286 genes were differentially expressed. Given the small size of the dataset, the data was primarily used as a whole for the purpose of pathway analysis. After correcting for multiple testing, none of the p-values for the identified genes remained significant.

Table 16. ***Most significantly up and downregulated genes in the CF-EXPR dataset.*** Genes are ranked based on moderated t-test non-adjusted p-values. P-values were adjusted using the Benjamini and Hochberg's (BH) method.

Gene symbol	Entrez gene ID	Fold change	p-value	Adjusted p-value[a]	Gene name[b]	Locus
Upregulated						
MEX3B	84206	1.26	4.25E-05	0.2130	mex-3 homolog B (C. elegans)	15q25.2
DNMT3B	1789	1.33	6.16E-05	0.2130	DNA (cytosine-5-)-methyltransferase 3 beta	20q11.2
TNF	7124	2.32	9.49E-05	0.2130	tumor necrosis factor	6p21.3
IFNG	3458	2.87	9.77E-05	0.2130	interferon, gamma	12q14
KIAA0495	57212	1.25	0.000197	0.3055	KIAA0495	1p36.32
IER5	51278	1.79	0.000260	0.3055	immediate early response 5	1q25.3
IGHV4-34	28395	1.41	0.000417	0.3530	immunoglobulin heavy variable 4-34	14q32.33
MYL4	4635	1.47	0.000887	0.4625	myosin, light chain 4, alkali; atrial, embryonic	17q21-qter
WDR62	284403	1.23	0.000973	0.4625	WD repeat domain 62	19q13.12
BGLAP	632	1.33	0.00148	0.4687	bone gamma-carboxyglutamate (gla) protein	1q25-q31
Downregulated						
GNPDA2	132789	-1.92	3.57E-06	0.0635	glucosamine-6-phosphate deaminase 2	4p13
WDR43	23160	-1.74	1.73E-05	0.1534	WD repeat domain 43	2p23.2
MRPL13	28998	-1.93	5.63E-05	0.2130	mitochondrial ribosomal protein L13	8q22.1-q22.3
DDX1	1653	-1.41	7.65E-05	0.2130	DEAD (Asp-Glu-Ala-Asp) box polypeptide 1	2p24
TP53RK	112858	-1.53	0.000108	0.2130	TP53 regulating kinase	20q13.2
C18orf19	125228	-1.38	0.000126	0.2240	chromosome 18 open reading frame 19	18p11.21
FAM175B	23172	-1.63	0.000226	0.3055	family with sequence similarity 175, member B	10q26.13
PPP1R8	5511	-1.61	0.000240	0.3055	protein phosphatase 1, regulatory (inhibitor) subunit 8	1p35
COMMD2	51122	-1.62	0.000279	0.3055	COMM domain containing 2	3q25.1
FBXO30	84085	-2.04	0.000280	0.3055	F-box protein 30	6q24

[a] Adjusted using the Benjamini and Hochberg's (BH) method.
[b] According to HUGO Gene Nomenclature Committee (HGNC).

5.3.5 Pathway analysis

GWAS datasets have only recently been used for other purposes than case-control association or CNV analysis. There is an increasing number of reports of pathway analysis using GWAS data (for e.g. Lesnick *et al.* 2007, Baranzini *et al.* 2009, Holmans *et al.* 2009, Wang *et al.* 2009b, O'Dushlaine *et al.* 2010), but the combined use of gene expression data and SNP data for the purpose of pathway analysis has mostly been limited to eQTL-based (expression quantitative trait locus) approaches (Zhong *et al.* 2010). Also, the pathway methodology has mainly focused on the analysis of the most significant association hits, which is likely to ignore a significant number of false negative hits, especially in phenotypes such as autism, where large effect size common variants cannot be expected.

In this study, we wanted to use pathway analysis to better address the role of common genetic variation in the CF-GWAS dataset. Since the dataset is small and has limited power to identify common variants on a genome-wide significant level, we wanted to include also the "grey zone" SNP results in the pathway analysis and try to partially overcome the problem of ignoring false negative results. We also wanted to extend the pathway analysis to global gene expression data from the same individuals to investigate whether changes observed at different levels of genomic data would reflect disturbances in same biological processes, thus providing an additional layer of confidence regarding the results. The key observations from the pathway analysis can be summarized as follows: i) apparent pathway level overlap was observed only between the two isolate datasets, CF-EXPR and CF-GWAS, which implicated vasculature development and axon guidance molecules in the pathogenesis of ASDs ii) overlap between the CF and the AGRE datasets was marginal, and iii) different p-value thresholds applied to the GWAS data did not markedly affect the pathway results. The ten most significant pathways from each dataset are presented in Tables 17-20.

A. Central Finland datasets

In both CF-GWAS and CF-EXPR datasets, multiple GO-categories related to vasculature development were identified among the 15 most significant pathways. These included processes such as angiogenesis, blood vessel development, and EGF signalling in the GWAS data, and vascular endothelial growth factor (VEGF) production-related categories in the expression data (Tables 17 and 18). Interestingly, angiogenesis and axon guidance are known to be partially regulated by same molecules (Adams and Eichmann 2010). All of these categories included genes, such as semaphorin 5A (*SEMA5A*) and neuropilin 2 (*NRP2*), which are known to function in axon guidance (Adams and Eichmann 2010). Hypoxia

inducible factor 1, alpha subunit (*HIF1A*) was the only overlapping gene in the vasculature-related pathways identified in the GWAS and gene expression datasets.

Other pathways which showed overlap between the two datasets were related to the actin cytoskeleton. CF-GWAS implicated "actin filament organization" whereas in CF-EXPR "actin cytoskeleton reorganization" was seen. Interestingly, actin cytoskeleton reorganization is closely associated with axon guidance through the axon growth cone. Other interesting categories were for example "protein polyubiquitination" in the CF-GWAS, a biological process implicated by a recent large CNV study in autism (Glessner *et al.* 2009), and two methyltransferase categories containing *DISC1*, which has been associated with ASDs in the same isolate (see Study I).

B. AGRE datasets

The most significant pathways observed in the CF datasets overlapped only occasionally with the pathways obtained from the AGRE datasets. Namely, helicase activity and protein ubiquitination were identified in both GWAS datasets, the latter of which has previously been linked to autism (Glessner *et al.* 2009). Further, the AGRE-GWAS implicated "positive regulation of neurogenesis" category with genes such as roundabout, axon guidance receptor, homolog 1 and 2 [Drosophila] (*ROBO1* and *ROBO2*), which link directly to vascular patterning and axon guidance, as identified in the CF data.

No apparently overlapping pathways between the AGRE-GWAS and the AGRE-EXPR datasets were seen (Tables 19 and 20). However, AGRE-GWAS had two categories related to nitric oxide metabolism and AGRE-EXPR yielded "regulation of glutamate signalling" and "postsynaptic density". Interestingly, nitric oxide is a known neurotransmitter in the brain, which is known to reinforce glutamatergic signalling as part of long-term potentiation (Haley *et al.* 1992).

Table 17. ***Summary of the pathway analysis results, Central Finland gene expression dataset (CF-EXPR).*** *Closely related pathways are counted as a single category. A minimum of two differentially expressed genes per pathway were required.*

Rank	GO ID	Pathway	Optimal p-value	Permuted p-value	Genes in pathway	Regulated genes
1	GO:0051640	organelle localization	6.78E-05	0.0006	35	19
2	GO:0000242	pericentriolar material	0.00016467	0.0007	3	2
3	GO:0004715	non-membrane spanning protein tyrosine kinase activity	6.26E-05	0.0008	20	8
4	GO:0031532	actin cytoskeleton reorganization	0.00017129	0.0011	8	4
5	GO:0001515	opioid peptide activity	0.00047084	0.0012	3	2
6	GO:0010573	vascular endothelial growth factor production	0.00033389	0.0021	6	5
7	GO:0043550	regulation of lipid kinase activity	0.00104429	0.0021	3	3
8	GO:0008156	negative regulation of DNA replication	0.00032367	0.0022	9	9
6	GO:0010574	regulation of vascular endothelial growth factor production	0.00033390	0.0022	6	5
9	GO:0034464	BBSome	0.00053425	0.0024	5	3
10	GO:0051640	regulation of hormone metabolic process	0.00051435	0.0026	8	7

ABBREVIATIONS: GO=Gene Ontology, BBS=Bardet-Biedl syndrome

Table 18. ***Summary of the pathway analysis results, Central Finland genome-wide association dataset (CF-GWAS).*** *Closely related pathways are counted as a single category. P-value threshold for individual marker significance in GWAS p=0.01. A minimum of two suggestively associated genes per pathway were required.*

Rank	GO ID	Pathway	Optimal p-value	Permuted p-value	Genes in pathway	Regulated genes
1	GO:0016528	sarcoplasm	1.69E-07	0.0001	26	6
2	GO:0032583	regulation of gene-specific transcription	0.00013053	0.0002	48	6
1	GO:0016529	sarcoplasmic reticulum	4.06E-06	0.0003	25	5
2	GO:0043193	positive regulation of gene-specific transcription	5.24E-05	0.0003	33	5
3	GO:0004700	atypical protein kinase C activity	0.00016893	0.0004	3	3
4	GO:0000775	chromosome, centromeric region	0.00338175	0.0007	77	3
4	GO:0000793	condensed chromosome	0.00390092	0.0015	81	3
5	GO:0001755	neural crest cell migration	0.00141258	0.0025	18	3
6	GO:0051124	synaptic growth at neuromuscular junction	0.00343003	0.0027	6	3
7	GO:0006110	regulation of glycolysis	0.00066297	0.003	11	3
8	GO:0001525	angiogenesis	1.92E-05	0.0031	145	10
8	GO:0001568	blood vessel development	5.77E-05	0.0031	199	11
9	GO:0008168	methyltransferase activity	0.0023585	0.0034	162	5
10	GO:0000209	protein polyubiquitination	0.03253891	0.0034	13	3

ABBREVIATIONS: GO=Gene Ontology

Table 19. ***Summary of the pathway analysis results, AGRE gene expression dataset (AGRE-EXPR).*** *Closely related pathways are counted as a single category. A minimum of two differentially expressed genes per pathway were required.*

Rank	GO ID	Pathway	Optimal p-value	Permuted p-value	Genes in pathway	Regulated genes
1	GO:0014069	postsynaptic density	0.00011479	0.0005	11	5
2	GO:0001539	ciliary or flagellar motility	0.00013554	0.0014	20	12
3	GO:0032297	negative regulation of DNA replication initiation	0.00095414	0.0022	3	2
4	GO:0005930	axoneme	0.00038024	0.0023	9	7
5	GO:0044447	axoneme part	0.00041054	0.0023	7	7
3	GO:0000076	DNA replication checkpoint	0.00095414	0.0023	3	2
5	GO:0005858	axonemal dynein complex	0.00041054	0.0026	7	7
3	GO:0030174	regulation of DNA replication initiation	0.00095414	0.0028	3	2
6	GO:0030159	receptor signaling complex scaffold activity	0.00073845	0.0029	4	4
7	GO:0006383	transcription from RNA polymerase III promoter	0.00066984	0.0031	8	7
8	GO:0004459	L-lactate dehydrogenase activity	0.00233629	0.0044	3	2
9	GO:0014048	regulation of glutamate secretion	0.00254639	0.0044	2	2
10	GO:0016783	sulfurtransferase activity	0.00262402	0.0051	2	2

ABBREVIATIONS: GO=Gene Ontology

Table 20. ***Summary of the pathway analysis results, AGRE genome-wide association dataset (AGRE-GWAS).*** *Closely related pathways are counted as a single category. P-value threshold for individual marker significance in GWAS p=0.01. A minimum of two suggestively associated genes per pathway were required.*

Rank	GO ID	Pathway	Optimal p-value	Permuted p-value	Genes in pathway	Regulated genes
1	GO:0005882	intermediate filament	3.07E-05	0.0001	129	5
1	GO:0045111	intermediate filament cytoskeleton	3.31E-05	0.0001	131	5
2	GO:0033177	proton-transporting two-sector ATPase complex, proton-transporting domain	0.00025466	0.0001	20	7
2	GO:0033179	proton-transporting V-type ATPase, V0 domain	2.84E-05	0.0002	8	5
1	GO:0045095	keratin filament	0.00047670	0.0002	63	3
3	GO:0004003	ATP-dependent DNA helicase activity	0.00012113	0.0003	23	3
3	GO:0003678	DNA helicase activity	0.00039583	0.0004	34	3
4	GO:0001636	corticotrophin-releasing factor gastric inhibitory peptide-like receptor activity	0.00139591	0.0008	5	3
5	GO:0035014	phosphoinositide 3-kinase regulator activity	0.00019815	0.0014	9	5
3	GO:0008026	ATP-dependent helicase activity	0.00078974	0.0015	98	4
3	GO:0070035	purine NTP-dependent helicase activity	0.00078974	0.0015	98	4
3	GO:0008094	DNA-dependent ATPase activity	0.00084989	0.0016	44	3
6	GO:0006809	nitric oxide biosynthetic process	0.00113630	0.0016	27	7
6	GO:0046209	nitric oxide metabolic process	0.00113630	0.0016	27	7
7	GO:0050901	leukocyte tethering or rolling	0.0048216	0.0017	5	3
2	GO:0033176	proton-transporting V-type ATPase complex	0.00085783	0.0018	21	6
8	GO:0042594	response to starvation	0.00387370	0.0025	30	8
3	GO:0004386	helicase activity	0.00162357	0.0039	138	5
9	GO:0006740	NADPH regeneration	0.00532658	0.0045	10	3
9	GO:0006098	pentose-phosphate shunt	0.00532658	0.0045	10	3
10	GO:0046323	glucose import	0.00162506	0.0058	16	4

ABBREVIATIONS: GO=Gene Ontology

C. Crohn's disease dataset

In order to further evaluate the performance of the GWANA method we applied it to a Crohn's disease (CD) GWAS dataset (1748 cases, 2938 controls) from the Wellcome Trust Case Control Consortium (WTCCC) (2007). Since the underlying biology of CD is better established than that of autism, certain biological processes could be expected to be found. Pathway analysis was performed using the same criteria and p-value thresholds as with the autism datasets.

Briefly, we identified many biological processes which are known to play a role in CD. These include "phagocytosis", "cellular response to unfolded protein", "defense response to Gram-negative bacteria", and "JAK-STAT cascade". Most of these were identified with all p-value thresholds, but the "JAK-STAT cascade" could be seen only with the broadest threshold. Notably, genes or pathways related to the HLA genes were not seen. Very similar results have been observed in a previous study, which performed a protein interaction network-based pathway analysis using the same CD dataset (Baranzini *et al.* 2009). In summary, applying the pathway analysis algorithm to the CD dataset demonstrated that GWANA can identify biologically relevant pathways from GWAS data, and provided an additional layer of confidence regarding the pathway results obtained with the ASD datasets.

D. Discussion on pathway analysis

The genetic background of CD is known to be different than that of autism, which may cause the pathway analysis to perform differently in these datasets. GWA studies in CD have demonstrated that a small group of risk alleles with relatively strong effects are present whereas in autism the hits are more widely spread and of lower risk. Since the GWANA method is specifically designed to analyze the "gray zone", and it operates with the ranking of markers or transcripts relative to each other rather than their individual significance to assess regulated pathways, it is likely to perform better with data with a relatively even distribution of moderately significant p-values. Thus, in fact, it should perform better with autism data than with CD data, and may even fail to recognize known CD pathways comprised of few, extremely significant hits. For example, the two genes most robustly associated to CD, *NOD2* and *IL23* (Barrett *et al.* 2008), do not appear among the regulated genes in the pathway analysis, and HLA genes were absent. This can also be due to the lack of evidence of interaction between these genes and the rest of the most significantly regulated genes. However, overall, the pathway analysis was able to identify multiple relevant pathways for CD, providing further support to the results obtained from the ASD datasets.

Since the pathway analysis method is based on the ranking of SNPs and transcripts relative to each other, rather than their individual significance, it is not clear how the significance of the obtained pathways should be evaluated, i.e. which pathways are more significant than others, and how much noise is introduced by small pathways which appear significant, possibly because of chance hits. This is clearly demonstrated by the Crohn's disease dataset, in which relevant processes were clearly identified, but which would be hard to separate from noise without prior knowledge of the biology. However, in the case of CD, all of the relevant pathways appeared to have a fairly large number of regulated genes, suggesting that filtering out the smallest categories might help to reduce false positives also in ASDs.

An obvious limitation to our pathway analysis method is that effects of variable LD are not considered or accounted for when mapping SNPs to their representative genes. The method was originally developed to be used with the Illumina HumanHap 300 series data, in which the genotyped SNPs are supposed to tag a single haplotype block, thereby reducing the problem of interdependent SNPs in pathway analysis. Also, the method uses the single most significant SNP to represent each gene, which, even though not ideal, is the most frequent approach taken by other studies as well (Baranzini *et al.* 2009). Thus, the statistic does not capture information of multiple distinct variants in a gene contributing to the overall association signal.

5.3.6 Conclusions

In this study, we made an effort to thoroughly investigate both common and rare genetic variation as well as gene expression and biological pathways in two extended ASD pedigrees from a population isolate. Our initial hypothesis was that the genealogical links would reflect IBD sharing, and that one or a few genomic regions shared by the affected individuals could be identified and used to pinpoint the causal variants. Given the well-known heterogeneity in ASDs, we did not expect to find haplotypes shared by all of the cases, but instead thought that a few haplotypes shared by ~half of the affected individuals would be present. Instead, the results obtained in Study III demonstrated that genetic heterogeneity in ASDs is so substantial, that even in genealogically connected individuals from a population isolate rare, family-specific variants probably comprise the majority of the overall genetic risk. We identified one genomic region in each pedigee where the majority of the affected individuals shared a homozygous segment of < 100 kb. Although small genomic regions can be shared just by chance, these regions should still be followed up as it is possible that they carry a shared risk variant.

We unexpectedly identified one genome-wide significant SNP at chromosome 2p16.1 in the GWA analysis of the CF-GWAS case-control dataset, which, together

with haplotype association analysis at this locus, implicates that there is an ASD risk variant present at this locus, in or near *BCL11A*. The function of *BCL11A*, together with convergent pathway analysis findings from SNP and gene expression data from the CF individuals, highlights axon guidance molecules in the pathogenesis of ASDs. As the most significant SNP association did not replicate in the nationwide Finnish family sample, the risk factor seems specific to the CF isolate.

Unless the increasing knowledge of the interplay of common and rare genetic variation reveals completely new and unexpected genetic mechanisms, it seems that the genetic background of autism is explained by rare, family specific genetic variants, and sequencing is the only method to tackle these. However, it is nevertheless possible that such scattered rare mutations, together with possible common variants, contribute to similar biological processes and pathways. In this study, we began exploring the biological processes behind ASDs using study samples from Central Finland and the AGRE. Pathway analysis of both GWAS and gene expression datasets from these study samples further supported family specificity, since the results obtained from the AGRE provided only little support for the pathways from the Finnish datasets. Clearly overlapping pathways related to vasculature development and reorganization of the actin cytoskeleton were observed in the CF-GWAS and CF-EXPR datasets. These identified pathways implicated multiple genes known to participate in axon guidance and are thus intriguing, given the obtained association finding to *BCL11A*. Also, synaptic cell-adhesion and cortical underconnectivity have recently been extensively discussed as potential biological causes for ASDs. Given the exploratory nature of the pathway analysis, the results should be considered suggestive only. Although isolate-specific results are generally challenging to replicate due to lack of independent, comparable datasets, validation of these pathways would provide valuable clues about the pathology underlying ASDs. However, if family-specific factors contribute to the disease susceptibility, replication might not be an appropriate method to validate initial findings. Parallel pathway analysis of SNP and gene expression data from the same study subjects might in part help to address this problem. To conclude, a comprehensive analysis of genealogically connected individuals with ASDs provided multiple lines of evidence for genes involved in axon guidance. We suggest that the genetic risk of ASDs in these individuals is likely to comprise of different combinations of rare genetic variants, and targeted resequencing of the identified regions of homozygosity, haplotypes, and linkage regions is essential to understand the underlying mutational spectrum.

6 CONCLUDING REMARKS AND FUTURE PROSPECTS

Human genetics has come a long way since the completion of the Human Genome Project. At the time, it took more than ten years, billions of dollars, and the effort of hundreds of scientists to sequence the haploid genome of one individual, whereas today, ten years later, it is entirely plausible and affordable to study millions of genomic loci in thousands of individuals in a timeline of months. The moment is quickly approaching when the full genomes of many individuals can be sequenced and analyzed routinely –examples of this are already emerging in the form of large-scale exome and whole genome sequencing initiatives, such as the 1000 Genomes project. Exciting years lie ahead for geneticists, when studying the entire genome at once – the very essence of genetics – will become routine.

Even though numerous new susceptibility variants for many disease phenotypes and other traits have been successfully discovered and replicated in GWA studies, linking single genetic variants to specific, and often biologically artificial, phenotypes has only covered a fraction of the genetic variation in these traits. This suggests that in most cases there is additional, rarer genetic variation contributing to the end-state phenotype. Also, the functional role of the most significantly associated SNPs from GWA studies has not been adequately defined, highlighting the need to study the effects the genetic variants have on various intermediate phenotypes at the cellular level, such as gene expression. Understanding how gene expression is regulated will therefore be key, and, at least in humans, will be challenging due to the astonishing variety of subtle regulatory mechanisms present, as discovered also in this project.

The pace of development and progress in methodology has been breathtaking: the last few years have seen a quick leap from genome-wide association studies to CNV and eQTL analysis, sequence-based transcriptomics, and analyses of epigenetic modifications. The long years of inconclusive linkage and candidate gene association studies have been mostly forgotten, and while it seems that many large-scale studies have essentially become hypothesis-free, the importance of targeted functional studies has become even more important. With genome-wide approaches providing detailed maps of where to look, there is still but a vague understanding of the mechanisms that mediate disease susceptibility conferred by a single variant. It has become evident that simply looking at the DNA sequence is not enough – additional layers of genomic data need to be added in order to elucidate the connections among sequence variants, transcripts, and eventually, functional protein products. The field will undoubtedly see many new "omics" in the years to come.

When this study was first started, knowledge of the genetic basis of autism spectrum disorders was on the same level as in most other psychiatric and neurological disorders. The era of GWA studies was yet to come, and it was generally accepted that these disorders have a complex genetic basis with multiple genetic factors, mostly common, increasing the risk of disease together with influence from environmental factors. The first large-scale GWA studies in ASDs published in 2009 yielded few common risk factors, whereas intriguing examples of rare, high-penetrance mutations in synaptic cell-adhesion genes started to emerge. It is now known that common variants do not play a significant role in ASDs, unless some previously unknown genetic mechanism masks their true effect from the current analysis methods available. Instead, evidence of family-specific rare genetic events such as mutations and *de novo* CNVs is accumulating, and it seems likely that a substantial proportion of ASD cases will be explained by these events, with or without interaction and influence with common polymorphisms.

The nature of rare genetic variants makes sequencing the superior method to find them. As next-generation sequencing methods become more widely available, a wave of exome sequencing studies in ASDs is expected in the coming years. Later, the approach is likely to extend to full genomes, since a proportion of the variants are likely to be non-coding, regulatory changes. Sequencing will be the next step in this study as well, since the affected individuals from the Central Finland pedigrees will soon undergo exome sequencing. Hopefully this will lead to the identification of the causal variants in these pedigrees.

It is often questioned whether the knowledge obtained from complex disease genetics studies will be of any use to the clinicians, early diagnosis, or the actual treatment of patients. Especially in the case of autism, where treatments currently seem to be light-years away, an earlier diagnosis might not be a significant improvement. Also, even if providing a molecular diagnosis is of great value to individual families, in terms of public health, identification of rare genetic causes for a disease is less exciting. Uniform applications, such as diagnostics, are hard to build on a scattered set of findings. However, even if "the autisms" present with multiple etiologies and painstaking heterogeneity, the strive to understand their origin is continuously teaching us novel things about the brain and neurobiology. Understanding autism would essentially mean understanding the brain, and this is, in fact, what should keep us going.

7 ACKNOWLEDGEMENTS

This study was carried out between years 2004 and 2010 at the National Public Health Institute (later National Institute for Health and Welfare; THL), Institute for Molecular Medicine Finland (FIMM), and Research Program of Molecular Neurology, University of Helsinki, Finland, as well as the Wellcome Trust Sanger Institute, Cambridge, UK. I wish to acknowledge Pekka Puska, director of THL, Anu Jalanko, head of the Public Health Genomics Unit, Olli Kallioniemi, director of FIMM, Anu Wartiovaara, head of the Molecular Neurology Program, and Mike Stratton, director of the Sanger Institute, for providing excellent research facilities.

This thesis was financially supported by the Research Foundation of the University of Helsinki, Biomedicum Helsinki Foundation, Orion-Farmos Research Foundation, Maud Kuistila Memorial Foundation, and University of Helsinki Medical Foundation, all of which are gratefully acknowledged.

My deepest gratitude goes to my supervisors, Professor Leena Palotie and Docent Iiris Hovatta. I feel privileged to have worked with such talented women, and I feel proud to represent the 3rd generation in the chain. Leena, who passed away nine months before my defense, taught me countless things about science and being a woman in science. Her endless drive and enthusiasm helped to overcome moments of despair and to believe in science. Thanks to her support, so many doors are now open for me, for which I am ever grateful. I feel extremely sad to have lost a unique and inspiring mentor, and I am sorry she could not see how well things turned out. The space she left behind will be very difficult to fill in so many ways. Iiris, your help, support, and rationality have been invaluable for me and this project. Your example as a scientist has made a huge impact on me and on the way I think science should be done. Your amazing ability to convince me that things will work out okay, even amidst complicated situations, kept me going. Thank you for always treating me more like a peer than a student, allowing me to find my own way.

A warm thank you also to Professor Aarno Palotie and Dr. Inês Barroso from the Sanger Institute, who looked after me and my projects after Leena's passing. Your help was very much appreciated.

A big thank you to Professor Mark McCarthy for accepting the role of the Opponent of my defense, despite his busy schedule. Docent Minna Männikkö and Professor Juha Partanen are acknowledged for valuable comments during the revision of this thesis, and Professor Olli Jänne for acting as the custos of the dissertation.

This study was made possible by the excellent, persistent work of our clinical collaborators, headed by Professor Lennart von-Wendt. I feel deeply sorry for his early passing, which is a huge loss for autism research in Finland. Taina Nieminen-von Wendt, Raija Vanhala, Reija Alen, and Susan Sarenius amongst many others are thanked for their efforts in sample collection and help in various other study sample-related matters. I am especially grateful to all of the Finnish families who participated to this study. Without your contribution none of this would have been possible. This study also includes data provided by the Autism Genetic Resource Exchange (AGRE) and the Wellcome Trust Case Control Consortium (WTCCC), both of which are acknowledged for making their valuable data available for the scientific community.

I wish to express my appreciation and gratitude to all of my collaborators and co-authors. Teppo Varilo is thanked for valuable help with genealogical data, as well as excellent comments on multiple manuscripts. Juha Saharinen is acknowledged for his ideas, efforts, and help with pathway analysis, and Dario Greco for help with gene expression and bioinformatics-related problems. Samuli Ripatti is thanked for assistance with various statistical issues. Further thanks to Mark Daly, Shaun Purcell, Eveliina Jakkula, Joni Turunen, Emilia Gaál, and Petri Auvinen for their contributions to the project. Docent Tiina Paunio, Dr. William Hennah, and Dr. Marika Palo are most kindly acknowledged for supervision, advice, and patience in the early days of this study, when I started in the lab as a second year biology student with no idea what a SNP was. A special thank you goes to Professor Dan Geschwind for his interest towards the project, help, support, and discussions about autism during his sabbatical in England, which I greatly enjoyed.

A huge thanks to Sari Kivikko and Chloe Noble, whose help over the years in a multitude of, sometimes unthinkable, matters has been invaluable. I also want to acknowledge the help of the secretaries, IT people, and lab technicians of the various units this work was carried out in.

I have been privileged to encounter and work with an amazing group of people, with whom I have shared some of the best years of my life. My research career would probably have turned out quite different, if it wasn't for the wonderful people in our autism project. The expertise and tireless support of my initial mentor and colleague, Tero Ylisaukko-oja, laid the foundation for my interest in, and my way of doing science, and gave me a kick-start to my thesis. Your influence is probably greater than you realize. Karola Rehnström is thanked for many years of close collaboration and friendship, which has been fruitful and enjoyable. Thank you for peer support during the recent times of uncertainty, as well as an unforgettable trip to Hawaii. I could not have wished for better colleagues.

My more recent colleagues, Mari Rossi and Mikko Muona are thanked for taking on the next generation of autism research in the group. I have no doubts both of you will do an excellent job. Elli Kempas, the unofficial guardian of our autism team throughout the years, is warmly thanked for all of her help and support. I wish you happy days of retirement.

I wish to thank all of my former and present-day colleagues from Leena's group and the rest of the department in Helsinki for a unique, inspirational, and supportive working environment: Annina, Annu, Ansku, Antti, Anu K, Anu L, Emma N, Emma P, Emmi, Hanski, Ida, Jarkko, Jenni, Juho, Jussi, Kaisa, Liisa, Marine, Mervi, Milja, Minna, Minttu, Nan, Pekka, Pia, PP, Suvi, Tiia, Tintti, Virpi, and the many others I may have forgotten. Thank you for the numerous crazy and not-so-crazy conversations, parties, unforgettable conference trips, and long days at the office - it has been a blast! A special thanks to Olli, with whom I will never get tired of discussing science and planning an infinite number of new projects. It's never a bad thing to aim high.

I also wish to acknowledge all of the people from the Research Program of Molecular Neurology. Although our encounters have been quite irregular and brief, I want to thank you of many fun moments and memories. Special thanks to Laura, Jonas, Pia, Tessa, Juuso and other members of group Hovatta for accepting me as a part-time member of the group.

A big thanks to all Sanger Human Genetics people, in particular Catherine Ingle, without whom my miRNA project would never be this far. My fellow Finns in Cambridge, Anu, Karola, Jonna, Eija, Katta, Verneri, Johannes, Kati, Henna, and Heidi, are thanked for making England feel less far away from home. Finally, huge thanks to Carl and Morven, Kate and Jules, Daniel and Ilana, Vesna, Don and Katinka for sharing a numerous wonderful trips and experiences, and making my time in Cambrigde so much fun. I will never forget the night of Sardines in the Wales-house.

My heartfelt thanks go to my dear old friends in Helsinki, Essi, Anna-Maija, Emma, and Ami. You have kept me sane during these years and made sure that I have not forgotten what the really important things in life are.

Finally, I am ever grateful to my dear family, Aarre, Leena, and Markku. Your unconditional love and support mean the world to me. And Jeff, thank you for helping me through these final, long months. You have opened my eyes to things I never imagined. Don't ever change.

Cambridge, November 2010
Helena Kilpinen

8 WEB-BASED RESOURCES

Patrocles	www.patrocles.org
PolymiRTS	http://compbio.uthsc.edu/miRSNP/home.php
miRBase	www.mirbase.org
TargetScan	www.targetscan.org
HapMap	www.hapmap.org
PFAM	http://pfam.sanger.ac.uk
UCSC	http://genome.ucsc.edu
Haploview	www.broadinstitute.org/haploview/haploview
HapMap	www.hapmap.org
Gene Ontology	www.geneontology.org
Ensembl	www.ensembl.org
OMIM	http://www.ncbi.nlm.nih.gov/omim
PLINK	http://pngu.mgh.harvard.edu/~purcell/plink
GEO	http://www.ncbi.nlm.nih.gov/geo
AGRE	www.agre.org
WTCCC	www.wtccc.org.uk
Autism Chromosome Rearrangement Database	http://projects.tcag.ca/autism
Decipher	https://decipher.sanger.ac.uk
EUCARUCA	http://agserver01.azn.nl:8080/ecaruca/ecaruca.jsp
Entrez	www.ncbi.nlm.nih.gov/Entrez
Brainarray	http://brainarray.mbni.med.umich.edu/Brainarray
NuGO R-server	http://nugo-r.bioinformatics.nl
R	www.r-project.org
Bioconductor	www.bioconductor.org
Flybase	http://flybase.org

9 REFERENCES

Abdulla, M. A., I. Ahmed, A. Assawamakin, J. Bhak, S. K. Brahmachari, G. C. Calacal, A. Chaurasia, C. H. Chen, J. Chen, Y. T. Chen, J. Chu, E. M. Cutiongco-de la Paz, M. C. De Ungria, F. C. Delfin, J. Edo, S. Fuchareon, H. Ghang, T. Gojobori, J. Han, S. F. Ho, B. P. Hoh, W. Huang, H. Inoko, P. Jha, T. A. Jinam, L. Jin, J. Jung, D. Kangwanpong, J. Kampuansai, G. C. Kennedy, P. Khurana, H. L. Kim, K. Kim, S. Kim, W. Y. Kim, K. Kimm, R. Kimura, T. Koike, S. Kulawonganunchai, V. Kumar, P. S. Lai, J. Y. Lee, S. Lee, E. T. Liu, P. P. Majumder, K. K. Mandapati, S. Marzuki, W. Mitchell, M. Mukerji, K. Naritomi, C. Ngamphiw, N. Niikawa, N. Nishida, B. Oh, S. Oh, J. Ohashi, A. Oka, R. Ong, C. D. Padilla, P. Palittapongarnpim, H. B. Perdigon, M. E. Phipps, E. Png, Y. Sakaki, J. M. Salvador, Y. Sandraling, V. Scaria, M. Seielstad, M. R. Sidek, A. Sinha, M. Srikummool, H. Sudoyo, S. Sugano, H. Suryadi, Y. Suzuki, K. A. Tabbada, A. Tan, K. Tokunaga, S. Tongsima, L. P. Villamor, E. Wang, Y. Wang, H. Wang, J. Y. Wu, H. Xiao, S. Xu, J. O. Yang, Y. Y. Shugart, H. S. Yoo, W. Yuan, G. Zhao and B. A. Zilfalil (2009). "Mapping human genetic diversity in Asia." *Science* **326**: 1541-5.

Abelson, J. F., K. Y. Kwan, B. J. O'Roak, D. Y. Baek, A. A. Stillman, T. M. Morgan, C. A. Mathews, D. L. Pauls, M. R. Rasin, M. Gunel, N. R. Davis, A. G. Ercan-Sencicek, D. H. Guez, J. A. Spertus, J. F. Leckman, L. S. t. Dure, R. Kurlan, H. S. Singer, D. L. Gilbert, A. Farhi, A. Louvi, R. P. Lifton, N. Sestan and M. W. State (2005). "Sequence variants in SLITRK1 are associated with Tourette's syndrome." *Science* **310**: 317-20.

Abrahams, B. S. and D. H. Geschwind (2008). "Advances in autism genetics: on the threshold of a new neurobiology." *Nat Rev Genet* **9**: 341-55.

Abu-Elneel, K., T. Liu, F. S. Gazzaniga, Y. Nishimura, D. P. Wall, D. H. Geschwind, K. Lao and K. S. Kosik (2008). "Heterogeneous dysregulation of microRNAs across the autism spectrum." *Neurogenetics* **9**: 153-61.

Adams, R. H. and A. Eichmann (2010). "Axon guidance molecules in vascular patterning." *Cold Spring Harb Perspect Biol* **2**: a001875.

Alarcon, M., B. S. Abrahams, J. L. Stone, J. A. Duvall, J. V. Perederiy, J. M. Bomar, J. Sebat, M. Wigler, C. L. Martin, D. H. Ledbetter, S. F. Nelson, R. M. Cantor and D. H. Geschwind (2008). "Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene." *Am J Hum Genet* **82**: 150-9.

Alarcon, M., R. M. Cantor, J. Liu, T. C. Gilliam and D. H. Geschwind (2002). "Evidence for a language quantitative trait locus on chromosome 7q in multiplex autism families." *Am J Hum Genet* **70**: 60-71.

Alarcon, M., A. L. Yonan, T. C. Gilliam, R. M. Cantor and D. H. Geschwind (2005). "Quantitative genome scan and Ordered-Subsets Analysis of autism endophenotypes support language QTLs." *Mol Psychiatry* **10**: 747-57.

Altmuller, J., L. J. Palmer, G. Fischer, H. Scherb and M. Wjst (2001). "Genomewide scans of complex human diseases: true linkage is hard to find." *Am J Hum Genet* **69**: 936-50.

Amaral, D. G., C. M. Schumann and C. W. Nordahl (2008). "Neuroanatomy of autism." *Trends Neurosci* **31**: 137-45.

American Psychiatric Association (1994). Diagnostic and Statistical Manual of Mental Disorders (4th edn) (DSM-IV). Washington, DC, APA.

Amir, R. E., I. B. Van den Veyver, M. Wan, C. Q. Tran, U. Francke and H. Y. Zoghbi (1999). "Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2." *Nat Genet* **23**: 185-8.

Anney, R., L. Klei, D. Pinto, R. Regan, J. Conroy, T. R. Magalhaes, C. Correia, B. S. Abrahams, N. Sykes, A. T. Pagnamenta, J. Almeida, E. Bacchelli, A. J. Bailey, G. Baird, A. Battaglia, T. Berney, N. Bolshakova, S. Bolte, P. F. Bolton, T. Bourgeron, S. Brennan, J. Brian, A. R. Carson, G. Casallo, J. Casey, S. H. Chu, L. Cochrane, C. Corsello, E. L. Crawford, A. Crossett, G. Dawson, M. de Jonge, R. Delorme, I. Drmic, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, E. Fombonne, C. M. Freitag, J. Gilbert, C. Gillberg, J. T. Glessner, J. Goldberg, J. Green, S. J. Guter, H. Hakonarson, E. A. Heron, M. Hill, R. Holt, J. L. Howe, G. Hughes, V. Hus, R. Igliozzi, C. Kim, S. M. Klauck, A. Kolevzon, O. Korvatska, V. Kustanovich, C. M. Lajonchere, J. A. Lamb, M. Laskawiec, M. Leboyer, A. Le Couteur, B. L. Leventhal, A. C. Lionel, X. Q. Liu, C. Lord, L. Lotspeich, S. C. Lund, E. Maestrini, W. Mahoney, C. Mantoulan, C. R. Marshall, H. McConachie, C. J. McDougle, J. McGrath, W. M. McMahon, N. M. Melhem, A. Merikangas, O. Migita, N. J. Minshew, G. K. Mirza, J. Munson, S. F. Nelson, C. Noakes, A. Noor, G. Nygren, G. Oliveira, K. Papanikolaou, J. R. Parr, B. Parrini, T. Paton, A. Pickles, J. Piven, D. J. Posey, A. Poustka, F. Poustka, A. Prasad, J. Ragoussis, K. Renshaw, J. Rickaby, W. Roberts, K. Roeder, B. Roge, M. L. Rutter, L. J. Bierut, J. P. Rice, J. Salt, K. Sansom, D. Sato, R. Segurado, L. Senman, N. Shah, V. C. Sheffield, L. Soorya, I. Sousa, V. Stoppioni, C. Strawbridge, R. Tancredi, K. Tansey, B. Thiruvahindrapduram, A. P. Thompson, S. Thomson, A. Tryfon, J. Tsiantis, H. Van Engeland, J. B. Vincent, F. Volkmar, S. Wallace, K. Wang, Z. Wang, T. H. Wassink, K. Wing, K. Wittemeyer, S. Wood, B. L. Yaspan, D. Zurawiecki, L. Zwaigenbaum, C. Betancur, J. D. Buxbaum, R. M. Cantor, E. H. Cook, H. Coon, M. L. Cuccaro, L. Gallagher, D. H. Geschwind, M. Gill, J. L. Haines, J. Miller, A. P. Monaco, J. I. Nurnberger, Jr., A. D. Paterson, M. A. Pericak-Vance, G. D. Schellenberg, S. W. Scherer, J. S. Sutcliffe, P. Szatmari, A. M. Vicente, V. J. Vieland, E. M. Wijsman, B. Devlin, S. Ennis and J. Hallmayer (2010). "A genome-wide scan for common alleles affecting risk for autism." *Hum Mol Genet*.

Arking, D. E., D. J. Cutler, C. W. Brune, T. M. Teslovich, K. West, M. Ikeda, A. Rea, M. Guy, S. Lin, E. H. Cook and A. Chakravarti (2008). "A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism." *Am J Hum Genet* **82**: 160-4.

Ashley-Koch, A., C. M. Wolpert, M. M. Menold, L. Zaeem, S. Basu, S. L. Donnelly, S. A. Ravan, C. M. Powell, M. B. Qumsiyeh, A. S. Aylsworth, J. M. Vance, J. R. Gilbert, H. H. Wright, R. K. Abramson, G. R. DeLong, M. L. Cuccaro and M. A. Pericak-Vance (1999). "Genetic studies of autistic disorder and chromosome 7." *Genomics* **61**: 227-36.

Asperger, H. (1944). "Die autistischen Psychopathen im Kindersalter." *Archiv für Psychiatrie und Nervenkrankheiten* **1**: 76-136.

Aulchenko, Y. S., S. Ripatti, I. Lindqvist, D. Boomsma, I. M. Heid, P. P. Pramstaller, B. W. Penninx, A. C. Janssens, J. F. Wilson, T. Spector, N. G. Martin, N. L. Pedersen, K. O. Kyvik, J. Kaprio, A. Hofman, N. B. Freimer, M. R. Jarvelin, U. Gyllensten, H. Campbell, I. Rudan, A. Johansson, F. Marroni, C. Hayward, V. Vitart, I. Jonasson, C. Pattaro, A. Wright, N. Hastie, I. Pichler, A. A. Hicks, M. Falchi, G. Willemsen, J. J. Hottenga, E. J. de Geus, G. W. Montgomery, J. Whitfield, P. Magnusson, J. Saharinen, M. Perola, K. Silander, A. Isaacs, E. J. Sijbrands, A. G. Uitterlinden, J. C. Witteman, B. A. Oostra, P. Elliott, A. Ruokonen, C. Sabatti, C. Gieger, T. Meitinger, F. Kronenberg, A. Doring, H. E. Wichmann, J. H. Smit, M. I. McCarthy, C. M. van Duijn and L. Peltonen (2009). "Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts." *Nat Genet* **41**: 47-55.

Auranen, M., R. Vanhala, T. Varilo, K. Ayers, E. Kempas, T. Ylisaukko-Oja, J. S. Sinsheimer, L. Peltonen and I. Jarvela (2002). "A genomewide screen for autism-spectrum disorders:

evidence for a major susceptibility locus on chromosome 3q25-27." *Am J Hum Genet* **71**: 777-90.

Auranen, M., T. Varilo, R. Alen, R. Vanhala, K. Ayers, E. Kempas, T. Ylisaukko-Oja, L. Peltonen and I. Jarvela (2003). "Evidence for allelic association on chromosome 3q25-27 in families with autism spectrum disorders originating from a subisolate of Finland." *Mol Psychiatry* **8**: 879-84.

Badner, J. A. and E. S. Gershon (2002a). "Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia." *Mol Psychiatry* **7**: 405-11.

Badner, J. A. and E. S. Gershon (2002b). "Regional meta-analysis of published data supports linkage of autism with markers on chromosome 7." *Mol Psychiatry* **7**: 56-66.

Bailey, A., A. Le Couteur, I. Gottesman, P. Bolton, E. Simonoff, E. Yuzda and M. Rutter (1995). "Autism as a strongly genetic disorder: evidence from a British twin study." *Psychol Med* **25**: 63-77.

Bailey, A., S. Palferman, L. Heavey and A. Le Couteur (1998). "Autism: the phenotype in relatives." *J Autism Dev Disord* **28**: 369-92.

Bailey, A., W. Phillips and M. Rutter (1996). "Autism: towards an integration of clinical, genetic, neuropsychological, and neurobiological perspectives." *J Child Psychol Psychiatry* **37**: 89-126.

Bakkaloglu, B., B. J. O'Roak, A. Louvi, A. R. Gupta, J. F. Abelson, T. M. Morgan, K. Chawarska, A. Klin, A. G. Ercan-Sencicek, A. A. Stillman, G. Tanriover, B. S. Abrahams, J. A. Duvall, E. M. Robbins, D. H. Geschwind, T. Biederer, M. Gunel, R. P. Lifton and M. W. State (2008). "Molecular cytogenetic analysis and resequencing of contactin associated protein-like 2 in autism spectrum disorders." *Am J Hum Genet* **82**: 165-73.

Bao, L., M. Zhou, L. Wu, L. Lu, D. Goldowitz, R. W. Williams and Y. Cui (2007). "PolymiRTS Database: linking polymorphisms in microRNA target sites with complex traits." *Nucleic Acids Res* **35**: D51-4.

Baranzini, S. E., N. W. Galwey, J. Wang, P. Khankhanian, R. Lindberg, D. Pelletier, W. Wu, B. M. Uitdehaag, L. Kappos, C. H. Polman, P. M. Matthews, S. L. Hauser, R. A. Gibson, J. R. Oksenberg and M. R. Barnes (2009). "Pathway and network-based analysis of genome-wide association studies in multiple sclerosis." *Hum Mol Genet* **18**: 2078-90.

Baron-Cohen, S. (2002). "The extreme male brain theory of autism." *Trends Cogn Sci* **6**: 248-254.

Baron-Cohen, S. and M. K. Belmonte (2005). "Autism: a window onto the development of the social and the analytic brain." *Annu Rev Neurosci* **28**: 109-26.

Baron-Cohen, S., R. C. Knickmeyer and M. K. Belmonte (2005). "Sex differences in the brain: implications for explaining autism." *Science* **310**: 819-23.

Baron, C. A., S. Y. Liu, C. Hicks and J. P. Gregg (2006a). "Utilization of lymphoblastoid cell lines as a system for the molecular modeling of autism." *J Autism Dev Disord* **36**: 973-82.

Baron, C. A., C. G. Tepper, S. Y. Liu, R. R. Davis, N. J. Wang, N. C. Schanen and J. P. Gregg (2006b). "Genomic and functional profiling of duplicated chromosome 15 cell lines reveal regulatory alterations in UBE3A-associated ubiquitin-proteasome pathway processes." *Hum Mol Genet* **15**: 853-69.

Barrett, J. C., B. Fry, J. Maller and M. J. Daly (2005). "Haploview: analysis and visualization of LD and haplotype maps." *Bioinformatics* **21**: 263-5.

Barrett, J. C., S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, J. D. Rioux, S. R. Brant, M. S. Silverberg, K. D. Taylor, M. M. Barmada, A. Bitton, T. Dassopoulos, L. W. Datta, T. Green, A. M. Griffiths, E. O. Kistner, M. T. Murtha, M. D. Regueiro, J. I. Rotter, L. P. Schumm, A. H. Steinhart, S. R. Targan, R. J. Xavier, C. Libioulle, C. Sandor, M. Lathrop, J. Belaiche, O. Dewit, I. Gut, S. Heath, D. Laukens, M. Mni, P. Rutgeerts, A. Van Gossum, D. Zelenika, D. Franchimont, J. P. Hugot, M. de Vos, S. Vermeire, E. Louis, L. R. Cardon, C. A. Anderson, H. Drummond, E. Nimmo, T. Ahmad, N. J.

Prescott, C. M. Onnie, S. A. Fisher, J. Marchini, J. Ghori, S. Bumpstead, R. Gwilliam, M. Tremelling, P. Deloukas, J. Mansfield, D. Jewell, J. Satsangi, C. G. Mathew, M. Parkes, M. Georges and M. J. Daly (2008). "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease." *Nat Genet* **40**: 955-62.

Barrett, S., J. C. Beck, R. Bernier, E. Bisson, T. A. Braun, T. L. Casavant, D. Childress, S. E. Folstein, M. Garcia, M. B. Gardiner, S. Gilman, J. L. Haines, K. Hopkins, R. Landa, N. H. Meyer, J. A. Mullane, D. Y. Nishimura, P. Palmer, J. Piven, J. Purdy, S. L. Santangelo, C. Searby, V. Sheffield, J. Singleton, S. Slager and et al. (1999). "An autosomal genomic screen for autism. Collaborative linkage study of autism." *Am J Med Genet* **88**: 609-15.

Bartel, D. P. (2009). "MicroRNAs: target recognition and regulatory functions." *Cell* **136**: 215-33.

Bartlett, C. W., R. Goedken and V. J. Vieland (2005). "Effects of updating linkage evidence across subsets of data: reanalysis of the autism genetic resource exchange data set." *Am J Hum Genet* **76**: 688-95.

Bauman, M. L. and T. L. Kemper (2005). "Neuroanatomic observations of the brain in autism: a review and future directions." *Int J Dev Neurosci* **23**: 183-7.

Bear, M. F., K. M. Huber and S. T. Warren (2004). "The mGluR theory of fragile X mental retardation." *Trends Neurosci* **27**: 370-7.

Behar, D. M., B. Yunusbayev, M. Metspalu, E. Metspalu, S. Rosset, J. Parik, S. Rootsi, G. Chaubey, I. Kutuev, G. Yudkovsky, E. K. Khusnutdinova, O. Balanovsky, O. Semino, L. Pereira, D. Comas, D. Gurwitz, B. Bonne-Tamir, T. Parfitt, M. F. Hammer, K. Skorecki and R. Villems (2010). "The genome-wide structure of the Jewish people." *Nature* **466**: 238-42.

Benayed, R., N. Gharani, I. Rossman, V. Mancuso, G. Lazar, S. Kamdar, S. E. Bruse, S. Tischfield, B. J. Smith, R. A. Zimmerman, E. Dicicco-Bloom, L. M. Brzustowicz and J. H. Millonig (2005). "Support for the homeobox transcription factor gene ENGRAILED 2 as an autism spectrum disorder susceptibility locus." *Am J Hum Genet* **77**: 851-68.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *J. Roy. Statist. Soc. Ser. B* **57**: 289-300.

Berkel, S., C. R. Marshall, B. Weiss, J. Howe, R. Roeth, U. Moog, V. Endris, W. Roberts, P. Szatmari, D. Pinto, M. Bonin, A. Riess, H. Engels, R. Sprengel, S. W. Scherer and G. A. Rappold (2010). "Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation." *Nat Genet* **42**: 489-91.

Bilguvar, K., A. K. Ozturk, A. Louvi, K. Y. Kwan, M. Choi, B. Tatli, D. Yalnizoglu, B. Tuysuz, A. O. Caglayan, S. Gokben, H. Kaymakcalan, T. Barak, M. Bakircioglu, K. Yasuno, W. Ho, S. Sanders, Y. Zhu, S. Yilmaz, A. Dincer, M. H. Johnson, R. A. Bronen, N. Kocer, H. Per, S. Mane, M. N. Pamir, C. Yalcinkaya, S. Kumandas, M. Topcu, M. Ozmen, N. Sestan, R. P. Lifton, M. W. State and M. Gunel (2010). "Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations." *Nature*.

Bittel, D. C., N. Kibiryeva and M. G. Butler (2007). "Whole genome microarray analysis of gene expression in subjects with fragile X syndrome." *Genet Med* **9**: 464-72.

Blackwood, D. H., A. Fordyce, M. T. Walker, D. M. St Clair, D. J. Porteous and W. J. Muir (2001). "Schizophrenia and affective disorders--cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family." *Am J Hum Genet* **69**: 428-33.

Blatt, G. J., C. M. Fitzgerald, J. T. Guptill, A. B. Booker, T. L. Kemper and M. L. Bauman (2001). "Density and distribution of hippocampal neurotransmitter receptors in autism: an autoradiographic study." *J Autism Dev Disord* **31**: 537-43.

Blomquist, H. K., M. Bohman, S. O. Edvinsson, C. Gillberg, K. H. Gustavson, G. Holmgren and J. Wahlstrom (1985). "Frequency of the fragile X syndrome in infantile autism. A Swedish multicenter study." *Clin Genet* **27**: 113-7.

Blouin, J. L., B. A. Dombroski, S. K. Nath, V. K. Lasseter, P. S. Wolyniec, G. Nestadt, M. Thornquist, G. Ullrich, J. McGrath, L. Kasch, M. Lamacz, M. G. Thomas, C. Gehrig, U. Radhakrishna, S. E. Snyder, K. G. Balk, K. Neufeld, K. L. Swartz, N. DeMarchi, G. N. Papadimitriou, D. G. Dikeos, C. N. Stefanis, A. Chakravarti, B. Childs, A. E. Pulver and et al. (1998). "Schizophrenia susceptibility loci on chromosomes 13q32 and 8p21." *Nat Genet* **20**: 70-3.

Bodmer, W. and C. Bonilla (2008). "Common and rare variants in multifactorial susceptibility to common diseases." *Nat Genet* **40**: 695-701.

Boucard, A. A., A. A. Chubykin, D. Comoletti, P. Taylor and T. C. Sudhof (2005). "A splice code for trans-synaptic cell adhesion mediated by binding of neuroligin 1 to alpha- and beta-neurexins." *Neuron* **48**: 229-36.

Bowman, E. P. (1988). "Asperger's syndrome and autism: the case for a connection." *Br J Psychiatry* **152**: 377-82.

Brandon, N. J., J. K. Millar, C. Korth, H. Sive, K. K. Singh and A. Sawa (2009). "Understanding the role of DISC1 in psychiatric disease and during normal development." *J Neurosci* **29**: 12768-75.

Brenman, J. E., J. R. Topinka, E. C. Cooper, A. W. McGee, J. Rosen, T. Milroy, H. J. Ralston and D. S. Bredt (1998). "Localization of postsynaptic density-93 to dendritic microtubules and interaction with microtubule-associated protein 1A." *J Neurosci* **18**: 8805-13.

Broman, K. W., J. C. Murray, V. C. Sheffield, R. L. White and J. L. Weber (1998). "Comprehensive human genetic maps: individual and sex-specific variation in recombination." *Am J Hum Genet* **63**: 861-9.

Brzustowicz, L. M., K. A. Hodgkinson, E. W. Chow, W. G. Honer and A. S. Bassett (2000). "Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-q22." *Science* **288**: 678-82.

Brzustowicz, L. M., W. G. Honer, E. W. Chow, D. Little, J. Hogan, K. Hodgkinson and A. S. Bassett (1999). "Linkage of familial schizophrenia to chromosome 13q32." *Am J Hum Genet* **65**: 1096-103.

Bucan, M., B. S. Abrahams, K. Wang, J. T. Glessner, E. I. Herman, L. I. Sonnenblick, A. I. Alvarez Retuerto, M. Imielinski, D. Hadley, J. P. Bradfield, C. Kim, N. B. Gidaya, I. Lindquist, T. Hutman, M. Sigman, V. Kustanovich, C. M. Lajonchere, A. Singleton, J. Kim, T. H. Wassink, W. M. McMahon, T. Owley, J. A. Sweeney, H. Coon, J. I. Nurnberger, M. Li, R. M. Cantor, N. J. Minshew, J. S. Sutcliffe, E. H. Cook, G. Dawson, J. D. Buxbaum, S. F. Grant, G. D. Schellenberg, D. H. Geschwind and H. Hakonarson (2009). "Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes." *PLoS Genet* **5**: e1000536.

Buono, R. J., F. W. Lohoff, T. Sander, M. R. Sperling, M. J. O'Connor, D. J. Dlugos, S. G. Ryan, G. T. Golden, H. Zhao, T. M. Scattergood, W. H. Berrettini and T. N. Ferraro (2004). "Association between variation in the human KCNJ10 potassium ion channel gene and seizure susceptibility." *Epilepsy Res* **58**: 175-83.

Burdick, K. E., C. A. Hodgkinson, P. R. Szeszko, T. Lencz, J. M. Ekholm, J. M. Kane, D. Goldman and A. K. Malhotra (2005). "DISC1 and neurocognitive function in schizophrenia." *Neuroreport* **16**: 1399-402.

Burdick, K. E., A. Kamiya, C. A. Hodgkinson, T. Lencz, P. DeRosse, K. Ishizuka, S. Elashvili, H. Arai, D. Goldman, A. Sawa and A. K. Malhotra (2008). "Elucidating the relationship between DISC1, NDEL1 and NDE1 and the risk for schizophrenia: evidence of epistasis and competitive binding." *Hum Mol Genet* **17**: 2462-73.

Buxbaum, J. D., J. M. Silverman, C. J. Smith, M. Kilifarski, J. Reichert, E. Hollander, B. A. Lawlor, M. Fitzgerald, D. A. Greenberg and K. L. Davis (2001). "Evidence for a susceptibility gene for autism on chromosome 2 and for genetic heterogeneity." *Am J Hum Genet* **68**: 1514-20.

Buyske, S. (2009). "Comment on the article "Heterogeneous dysregulation of microRNAs across the autism spectrum" by Abu-Elneel et al." *Neurogenetics* **10**: 167; author reply 169-70.

Callicott, J. H., R. E. Straub, L. Pezawas, M. F. Egan, V. S. Mattay, A. R. Hariri, B. A. Verchinski, A. Meyer-Lindenberg, R. Balkissoon, B. Kolachana, T. E. Goldberg and D. R. Weinberger (2005). "Variation in DISC1 affects hippocampal structure and function and increases risk for schizophrenia." *Proc Natl Acad Sci U S A* **102**: 8627-32.

Camargo, L. M., V. Collura, J. C. Rain, K. Mizuguchi, H. Hermjakob, S. Kerrien, T. P. Bonnert, P. J. Whiting and N. J. Brandon (2007). "Disrupted in Schizophrenia 1 Interactome: evidence for the close connectivity of risk genes and a potential synaptic basis for schizophrenia." *Mol Psychiatry* **12**: 74-86.

Campbell, D. B., R. D'Oronzio, K. Garbett, P. J. Ebert, K. Mirnics, P. Levitt and A. M. Persico (2007). "Disruption of cerebral cortex MET signaling in autism spectrum disorder." *Ann Neurol* **62**: 243-50.

Campbell, D. B., J. S. Sutcliffe, P. J. Ebert, R. Militerni, C. Bravaccio, S. Trillo, M. Elia, C. Schneider, R. Melmed, R. Sacco, A. M. Persico and P. Levitt (2006). "A genetic variant that disrupts MET transcription is associated with autism." *Proc Natl Acad Sci U S A* **103**: 16834-9.

Cannon, T. D., W. Hennah, T. G. van Erp, P. M. Thompson, J. Lonnqvist, M. Huttunen, T. Gasperoni, A. Tuulio-Henriksson, T. Pirkola, A. W. Toga, J. Kaprio, J. Mazziotta and L. Peltonen (2005). "Association of DISC1/TRAX haplotypes with schizophrenia, reduced prefrontal gray matter, and impaired short- and long-term memory." *Arch Gen Psychiatry* **62**: 1205-13.

Cantor, R. M., N. Kono, J. A. Duvall, A. Alvarez-Retuerto, J. L. Stone, M. Alarcon, S. F. Nelson and D. H. Geschwind (2005). "Replication of autism linkage: fine-mapping peak at 17q21." *Am J Hum Genet* **76**: 1050-6.

Cardon, L. R. and J. I. Bell (2001). "Association study designs for complex diseases." *Nat Rev Genet* **2**: 91-9.

Chakrabarti, S. and E. Fombonne (2001). "Pervasive developmental disorders in preschool children." *Jama* **285**: 3093-9.

Chakrabarti, S. and E. Fombonne (2005). "Pervasive developmental disorders in preschool children: confirmation of high prevalence." *Am J Psychiatry* **162**: 1133-41.

Chakravarti, A. (1999). "Population genetics--making sense out of sequence." *Nat Genet* **21**: 56-60.

Charlesworth, J. C., J. E. Curran, M. P. Johnson, H. H. Goring, T. D. Dyer, V. P. Diego, J. W. Kent, Jr., M. C. Mahaney, L. Almasy, J. W. MacCluer, E. K. Moses and J. Blangero (2010). "Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes." *BMC Med Genomics* **3**: 29.

Chen, G. and A. J. Courey (2000). "Groucho/TLE family proteins and transcriptional repression." *Gene* **249**: 1-16.

Chubb, J. E., N. J. Bradshaw, D. C. Soares, D. J. Porteous and J. K. Millar (2008). "The DISC locus in psychiatric illness." *Mol Psychiatry* **13**: 36-64.

Chugani, D. C. (2004). "Serotonin in autism and pediatric epilepsies." *Ment Retard Dev Disabil Res Rev* **10**: 112-6.

Clayton, D. (1999). "A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission." *Am J Hum Genet* **65**: 1170-7.

Clop, A., F. Marcq, H. Takeda, D. Pirottin, X. Tordoir, B. Bibe, J. Bouix, F. Caiment, J. M. Elsen, F. Eychenne, C. Larzul, E. Laville, F. Meish, D. Milenkovic, J. Tobin, C. Charlier and M. Georges (2006). "A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep." *Nat Genet* **38**: 813-8.

Cody, H., K. Pelphrey and J. Piven (2002). "Structural and functional magnetic resonance imaging of autism." *Int J Dev Neurosci* **20**: 421-38.

Cole, K. A., D. B. Krizman and M. R. Emmert-Buck (1999). "The genetics of cancer--a 3D model." *Nat Genet* **21**: 38-41.

Conrad, D. F., T. D. Andrews, N. P. Carter, M. E. Hurles and J. K. Pritchard (2006). "A high-resolution survey of deletion polymorphism in the human genome." *Nat Genet* **38**: 75-81.

Conrad, D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. Macarthur, J. R. Macdonald, I. Onyiah, A. W. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer and M. E. Hurles (2010). "Origins and functional impact of copy number variation in the human genome." *Nature* **464**: 704-12.

Coon, H., N. Matsunami, J. Stevens, J. Miller, C. Pingree, N. J. Camp, A. Thomas, L. Krasny, J. Lainhart, M. F. Leppert and W. McMahon (2005). "Evidence for linkage on chromosome 3q25-27 in a large autism extended pedigree." *Hum Hered* **60**: 220-6.

Corder, E. H., A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, G. W. Small, A. D. Roses, J. L. Haines and M. A. Pericak-Vance (1993). "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families." *Science* **261**: 921-3.

Courchesne, E. and K. Pierce (2005). "Why the frontal cortex in autism might be talking only to itself: local over-connectivity but long-distance disconnection." *Curr Opin Neurobiol* **15**: 225-30.

Dai, M., P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson and F. Meng (2005). "Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data." *Nucleic Acids Res* **33**: e175.

Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson and E. S. Lander (2001). "High-resolution haplotype structure in the human genome." *Nat Genet* **29**: 229-32.

de Bakker, P. I., R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly and D. Altshuler (2005). "Efficiency and power in genetic association studies." *Nat Genet* **37**: 1217-23.

De La Vega, F. M., H. Isaac, A. Collins, C. R. Scafe, B. V. Halldorsson, X. Su, R. A. Lippert, Y. Wang, M. Laig-Webster, R. T. Koehler, J. S. Ziegle, L. T. Wogan, J. F. Stevens, K. M. Leinen, S. J. Olson, K. J. Guegler, X. You, L. H. Xu, H. G. Hemken, F. Kalush, M. Itakura, Y. Zheng, G. de The, S. J. O'Brien, A. G. Clark, S. Istrail, M. W. Hunkapiller, E. G. Spier and D. A. Gilbert (2005). "The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern." *Genome Res* **15**: 454-62.

DeRosse, P., C. A. Hodgkinson, T. Lencz, K. E. Burdick, J. M. Kane, D. Goldman and A. K. Malhotra (2007). "Disrupted in schizophrenia 1 genotype and positive symptoms in schizophrenia." *Biol Psychiatry* **61**: 1208-10.

Dimas, A. S., S. Deutsch, B. E. Stranger, S. B. Montgomery, C. Borel, H. Attar-Cohen, C. Ingle, C. Beazley, M. Gutierrez Arcelus, M. Sekowska, M. Gagnebin, J. Nisbett, P. Deloukas, E. T. Dermitzakis and S. E. Antonarakis (2009). "Common regulatory variation impacts gene expression in a cell type-dependent manner." *Science* **325**: 1246-50.

Dixon, A. L., L. Liang, M. F. Moffatt, W. Chen, S. Heath, K. C. Wong, J. Taylor, E. Burnett, I. Gut, M. Farrall, G. M. Lathrop, G. R. Abecasis and W. O. Cookson (2007). "A genome-wide association study of global gene expression." *Nat Genet* **39**: 1202-7.

Doench, J. G., C. P. Petersen and P. A. Sharp (2003). "siRNAs can function as miRNAs." *Genes Dev* **17**: 438-42.

Duan, X., J. H. Chang, S. Ge, R. L. Faulkner, J. Y. Kim, Y. Kitabatake, X. B. Liu, C. H. Yang, J. D. Jordan, D. K. Ma, C. Y. Liu, S. Ganesan, H. J. Cheng, G. L. Ming, B. Lu and H. Song (2007). "Disrupted-In-Schizophrenia 1 regulates integration of newly generated neurons in the adult brain." *Cell* **130**: 1146-58.

Duvall, J. A., A. Lu, R. M. Cantor, R. D. Todd, J. N. Constantino and D. H. Geschwind (2007). "A quantitative trait locus analysis of social responsiveness in multiplex autism families." *Am J Psychiatry* **164**: 656-62.

Ehlers, S. and C. Gillberg (1993). "The epidemiology of Asperger syndrome. A total population study." *J Child Psychol Psychiatry* **34**: 1327-50.

Ehlers, S., C. Gillberg and L. Wing (1999). "A screening questionnaire for Asperger syndrome and other high- functioning autism spectrum disorders in school age children." *J Autism Dev Disord* **29**: 129-41.

Ekelund, J., W. Hennah, T. Hiekkalinna, A. Parker, J. Meyer, J. Lonnqvist and L. Peltonen (2004). "Replication of 1q42 linkage in Finnish schizophrenia pedigrees." *Mol Psychiatry* **9**: 1037-41.

Ekelund, J., I. Hovatta, A. Parker, T. Paunio, T. Varilo, R. Martin, J. Suhonen, P. Ellonen, G. Chan, J. S. Sinsheimer, E. Sobel, H. Juvonen, R. Arajarvi, T. Partonen, J. Suvisaari, J. Lonnqvist, J. Meyer and L. Peltonen (2001). "Chromosome 1 loci in Finnish schizophrenia families." *Hum Mol Genet* **10**: 1611-7.

Ellegren, H. (2000). "Microsatellite mutations in the germline: implications for evolutionary inference." *Trends Genet* **16**: 551-8.

Fabian, M. R., N. Sonenberg and W. Filipowicz (2010). "Regulation of mRNA translation and stability by microRNAs." *Annu Rev Biochem* **79**: 351-79.

Fatemi, S. H. (2002). "The role of Reelin in pathology of autism." *Mol Psychiatry* **7**: 919-20.

Fatemi, S. H. (2004). "Reelin glycoprotein: structure, biology and roles in health and disease." *Mol Psychiatry*.

Feng, J., R. Schroer, J. Yan, W. Song, C. Yang, A. Bockholt, E. H. Cook, Jr., C. Skinner, C. E. Schwartz and S. S. Sommer (2006). "High frequency of neurexin 1beta signal peptide structural variants in patients with autism." *Neurosci Lett* **409**: 10-3.

Ferraro, T. N., G. T. Golden, G. G. Smith, J. F. Martin, F. W. Lohoff, T. A. Gieringer, D. Zamboni, C. L. Schwebel, D. M. Press, S. O. Kratzer, H. Zhao, W. H. Berrettini and R. J. Buono (2004). "Fine mapping of a seizure susceptibility locus on mouse Chromosome 1: nomination of Kcnj10 as a causative gene." *Mamm Genome* **15**: 239-51.

Feuk, L., A. R. Carson and S. W. Scherer (2006). "Structural variation in the human genome." *Nat Rev Genet* **7**: 85-97.

Firth, H. V., S. M. Richards, A. P. Bevan, S. Clayton, M. Corpas, D. Rajan, S. Van Vooren, Y. Moreau, R. M. Pettett and N. P. Carter (2009). "DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources." *Am J Hum Genet* **84**: 524-33.

Folstein, S. and M. Rutter (1977). "Infantile autism: a genetic study of 21 twin pairs." *J Child Psychol Psychiatry* **18**: 297-321.

Folstein, S. E. and B. Rosen-Sheidley (2001). "Genetics of autism: complex aetiology for a heterogeneous disorder." *Nat Rev Genet* **2**: 943-55.

Fombonne, E. (1999). "The epidemiology of autism: a review." *Psychol Med* **29**: 769-86.

Fombonne, E. (2005). "Epidemiology of autistic disorder and other pervasive developmental disorders." *J Clin Psychiatry* **66 Suppl 10**: 3-8.

Fombonne, E. (2009). "Epidemiology of pervasive developmental disorders." *Pediatr Res* **65**: 591-8.

Fombonne, E., B. Roge, J. Claverie, S. Courty and J. Fremolle (1999). "Microcephaly and macrocephaly in autism." *J Autism Dev Disord* **29**: 113-9.

Frazer, K. A., D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archeveque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson and J. Stewart (2007). "A second generation human haplotype map of over 3.1 million SNPs." *Nature* **449**: 851-61.

Freitag, C. M. (2007). "The genetics of autistic disorders and its clinical relevance: a review of the literature." *Mol Psychiatry* **12**: 2-22.

Friedman, R. C., K. K. Farh, C. B. Burge and D. P. Bartel (2009). "Most mammalian mRNAs are conserved targets of microRNAs." *Genome Res* **19**: 92-105.

Frith, C. (2004). "Is autism a disconnection disorder?" *Lancet Neurol* **3**: 577.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R.

Cooper, R. Ward, E. S. Lander, M. J. Daly and D. Altshuler (2002). "The structure of haplotype blocks in the human genome." *Science* **296**: 2225-9.

Gautier, L., L. Cope, B. M. Bolstad and R. A. Irizarry (2004). "affy--analysis of Affymetrix GeneChip data at the probe level." *Bioinformatics* **20**: 307-15.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang and J. Zhang (2004). "Bioconductor: open software development for computational biology and bioinformatics." *Genome Biol* **5**: R80.

Geschwind, D. H. and P. Levitt (2007). "Autism spectrum disorders: developmental disconnection syndromes." *Curr Opin Neurobiol* **17**: 103-11.

Geschwind, D. H., J. Sowinski, C. Lord, P. Iversen, J. Shestack, P. Jones, L. Ducat and S. J. Spence (2001). "The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions." *Am J Hum Genet* **69**: 463-6.

Gharani, N., R. Benayed, V. Mancuso, L. M. Brzustowicz and J. H. Millonig (2004). "Association of the homeobox transcription factor, ENGRAILED 2, 3, with autism spectrum disorder." *Mol Psychiatry* **9**: 474-84.

Gillberg, C. (1989). "Asperger syndrome in 23 Swedish children." *Dev Med Child Neurol* **31**: 520-31.

Gillberg, C. (1998). "Chromosomal disorders and autism." *J Autism Dev Disord* **28**: 415-25.

Gillberg, C. and E. Billstedt (2000). "Autism and Asperger syndrome: coexistence with other clinical disorders." *Acta Psychiatr Scand* **102**: 321-30.

Gillberg, C., M. Rastam and E. Wentz (2001). "The Asperger Syndrome (and high-functioning autism) Diagnostic Interview (ASDI): a preliminary study of a new structured clinical interview." *Autism* **5**: 57-66.

Gillberg, I. C. and C. Gillberg (1989). "Asperger syndrome--some epidemiological considerations: a research note." *J Child Psychol Psychiatry* **30**: 631-8.

Glessner, J. T., K. Wang, G. Cai, O. Korvatska, C. E. Kim, S. Wood, H. Zhang, A. Estes, C. W. Brune, J. P. Bradfield, M. Imielinski, E. C. Frackelton, J. Reichert, E. L. Crawford, J. Munson, P. M. Sleiman, R. Chiavacci, K. Annaiah, K. Thomas, C. Hou, W. Glaberson, J. Flory, F. Otieno, M. Garris, L. Soorya, L. Klei, J. Piven, K. J. Meyer, E. Anagnostou, T. Sakurai, R. M. Game, D. S. Rudd, D. Zurawiecki, C. J. McDougle, L. K. Davis, J. Miller, D. J. Posey, S. Michaels, A. Kolevzon, J. M. Silverman, R. Bernier, S. E. Levy, R. T. Schultz, G. Dawson, T. Owley, W. M. McMahon, T. H. Wassink, J. A. Sweeney, J. I. Nurnberger, H. Coon, J. S. Sutcliffe, N. J. Minshew, S. F. Grant, M. Bucan, E. H. Cook, J. D. Buxbaum, B. Devlin, G. D. Schellenberg and H. Hakonarson (2009). "Autism genome-wide copy number variation reveals ubiquitin and neuronal genes." *Nature* **459**: 569-73.

Goring, H. H., J. E. Curran, M. P. Johnson, T. D. Dyer, J. Charlesworth, S. A. Cole, J. B. Jowett, L. J. Abraham, D. L. Rainwater, A. G. Comuzzie, M. C. Mahaney, L. Almasy, J. W. MacCluer, A. H. Kissebah, G. R. Collier, E. K. Moses and J. Blangero (2007). "Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes." *Nat Genet* **39**: 1208-16.

Göring, H. H. and J. D. Terwilliger (2000). "Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified." *Am J Hum Genet* **66**: 1310-27.

Gregg, J. P., L. Lit, C. A. Baron, I. Hertz-Picciotto, W. Walker, R. A. Davis, L. A. Croen, S. Ozonoff, R. Hansen, I. N. Pessah and F. R. Sharp (2008). "Gene expression changes in children with autism." *Genomics* **91**: 22-9.

Griffiths-Jones, S., R. J. Grocock, S. van Dongen, A. Bateman and A. J. Enright (2006). "miRBase: microRNA sequences, targets and gene nomenclature." *Nucleic Acids Res* **34**: D140-4.

Grimson, A., K. K. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim and D. P. Bartel (2007). "MicroRNA targeting specificity in mammals: determinants beyond seed pairing." *Mol Cell* **27**: 91-105.

Guo, H., N. T. Ingolia, J. S. Weissman and D. P. Bartel (2010). "Mammalian microRNAs predominantly act to decrease target mRNA levels." *Nature* **466**: 835-40.

Gupta, A. R. and M. W. State (2007). "Recent advances in the genetics of autism." *Biol Psychiatry* **61**: 429-37.

Gurling, H. M., G. Kalsi, J. Brynjolfson, T. Sigmundsson, R. Sherrington, B. S. Mankoo, T. Read, P. Murphy, E. Blaveri, A. McQuillin, H. Petursson and D. Curtis (2001). "Genomewide genetic linkage analysis confirms the presence of susceptibility loci for schizophrenia, on chromosomes 1q32.2, 5q33.2, and 8p21-22 and provides support for linkage to schizophrenia, on chromosomes 11q23.3-24 and 20q12.1-11.23." *Am J Hum Genet* **68**: 661-73.

Haley, J. E., G. L. Wilcox and P. F. Chapman (1992). "The role of nitric oxide in hippocampal long-term potentiation." *Neuron* **8**: 211-6.

Happe, F., R. Booth, R. Charlton and C. Hughes (2006). "Executive function deficits in autism spectrum disorders and attention-deficit/hyperactivity disorder: examining profiles across domains and ages." *Brain Cogn* **61**: 25-39.

Hashimoto, R., T. Numakawa, T. Ohnishi, E. Kumamaru, Y. Yagasaki, T. Ishimoto, T. Mori, K. Nemoto, N. Adachi, A. Izumi, S. Chiba, H. Noguchi, T. Suzuki, N. Iwata, N. Ozaki, T. Taguchi, A. Kamiya, A. Kosuga, M. Tatsumi, K. Kamijima, D. R. Weinberger, A. Sawa and H. Kunugi (2006). "Impact of the DISC1 Ser704Cys polymorphism on risk for major depression, brain morphology and ERK signaling." *Hum Mol Genet* **15**: 3024-33.

Heitzler, P., M. Bourouis, L. Ruel, C. Carteret and P. Simpson (1996). "Genes of the Enhancer of split and achaete-scute complexes are required for a regulatory loop between Notch and Delta during lateral signalling in Drosophila." *Development* **122**: 161-71.

Hennah, W., P. Thomson, L. Peltonen and D. Porteous (2006). "Genes and schizophrenia: beyond schizophrenia: the role of DISC1 in major mental illness." *Schizophr Bull* **32**: 409-16.

Hennah, W., A. Tuulio-Henriksson, T. Paunio, J. Ekelund, T. Varilo, T. Partonen, T. D. Cannon, J. Lonnqvist and L. Peltonen (2005). "A haplotype within the DISC1 gene is associated with visual memory functions in families with a high density of schizophrenia." *Mol Psychiatry* **10**: 1097-103.

Hennah, W., T. Varilo, M. Kestila, T. Paunio, R. Arajarvi, J. Haukka, A. Parker, R. Martin, S. Levitzky, T. Partonen, J. Meyer, J. Lonnqvist, L. Peltonen and J. Ekelund (2003). "Haplotype transmission analysis provides evidence of association for DISC1 to schizophrenia and suggests sex-dependent effects." *Hum Mol Genet* **12**: 3151-9.

Hiard, S., C. Charlier, W. Coppieters, M. Georges and D. Baurain (2010). "Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates." *Nucleic Acids Res* **38**: D640-51.

Hodgkinson, C. A., D. Goldman, J. Jaeger, S. Persaud, J. M. Kane, R. H. Lipsky and A. K. Malhotra (2004). "Disrupted in schizophrenia 1 (DISC1): association with schizophrenia, schizoaffective disorder, and bipolar disorder." *Am J Hum Genet* **75**: 862-72.

Hollander, J. A., H. I. Im, A. L. Amelio, J. Kocerha, P. Bali, Q. Lu, D. Willoughby, C. Wahlestedt, M. D. Conkright and P. J. Kenny (2010). "Striatal microRNA controls cocaine intake through CREB signalling." *Nature* **466**: 197-202.

Holmans, P., E. K. Green, J. S. Pahwa, M. A. Ferreira, S. M. Purcell, P. Sklar, M. J. Owen, M. C. O'Donovan and N. Craddock (2009). "Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder." *Am J Hum Genet* **85**: 13-24.

Hong, E. J., A. E. West and M. E. Greenberg (2005). "Transcriptional control of cognitive development." *Curr Opin Neurobiol* **15**: 21-8.

Horvath, S., X. Xu and N. M. Laird (2001). "The family based association test method: strategies for studying general genotype--phenotype associations." *Eur J Hum Genet* **9**: 301-6.

Hu, V. W., A. Nguyen, K. S. Kim, M. E. Steinberg, T. Sarachana, M. A. Scully, S. J. Soldin, T. Luu and N. H. Lee (2009a). "Gene expression profiling of lymphoblasts from autistic and nonaffected sib pairs: altered pathways in neuronal development and steroid biosynthesis." *PLoS One* **4**: e5775.

Hu, V. W., T. Sarachana, K. S. Kim, A. Nguyen, S. Kulkarni, M. E. Steinberg, T. Luu, Y. Lai and N. H. Lee (2009b). "Gene expression profiling differentiates autism case-controls and phenotypic variants of autism spectrum disorders: evidence for circadian rhythm dysfunction in severe autism." *Autism Res* **2**: 78-97.

Hu, V. W. and M. E. Steinberg (2009). "Novel clustering of items from the Autism Diagnostic Interview-Revised to define phenotypes within autism spectrum disorders." *Autism Res* **2**: 67-77.

Hughes, J. R. (2007). "Autism: the first firm finding = underconnectivity?" *Epilepsy Behav* **11**: 20-4.

Iafrate, A. J., L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer and C. Lee (2004). "Detection of large-scale variation in the human genome." *Nat Genet* **36**: 949-51.

Ikuta, J., A. Maturana, T. Fujita, T. Okajima, K. Tatematsu, K. Tanizawa and S. Kuroda (2007). "Fasciculation and elongation protein zeta-1 (FEZ1) participates in the polarization of hippocampal neuron by controlling the mitochondrial motility." *Biochem Biophys Res Commun* **353**: 127-32.

IMGSAC (1998). "A full genome screen for autism with evidence for linkage to a region on chromosome 7q. International Molecular Genetic Study of Autism Consortium." *Hum Mol Genet* **7**: 571-8.

IMGSAC (2001a). "Further characterization of the autism susceptibility locus AUTS1 on chromosome 7q." *Hum Mol Genet* **10**: 973-82.

IMGSAC (2001b). "A genomewide screen for autism: strong evidence for linkage to chromosomes 2q, 7q, and 16p." *Am J Hum Genet* **69**: 570-81.

Irie, M., Y. Hata, M. Takeuchi, K. Ichtchenko, A. Toyoda, K. Hirao, Y. Takai, T. W. Rosahl and T. C. Sudhof (1997). "Binding of neuroligins to PSD-95." *Science* **277**: 1511-5.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." *Biostatistics* **4**: 249-64.

Jakkula, E., V. Leppa, A. M. Sulonen, T. Varilo, S. Kallio, A. Kemppinen, S. Purcell, K. Koivisto, P. Tienari, M. L. Sumelahti, I. Elovaara, T. Pirttila, M. Reunanen, A. Aromaa, A. B. Oturai, H. B. Sondergaard, H. F. Harbo, I. L. Mero, S. B. Gabriel, D. B. Mirel, S. L. Hauser, L. Kappos, C. Polman, P. L. De Jager, D. A. Hafler, M. J. Daly, A. Palotie, J. Saarela and L. Peltonen (2010). "Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in STAT3 gene." *Am J Hum Genet* **86**: 285-91.

Jakkula, E., K. Rehnstrom, T. Varilo, O. P. Pietilainen, T. Paunio, N. L. Pedersen, U. deFaire, M. R. Jarvelin, J. Saharinen, N. Freimer, S. Ripatti, S. Purcell, A. Collins, M. J. Daly, A. Palotie and L. Peltonen (2008). "The genome-wide patterns of variation expose significant substructure in a founder population." *Am J Hum Genet* **83**: 787-94.

Jamain, S., H. Quach, C. Betancur, M. Rastam, C. Colineaux, I. C. Gillberg, H. Soderstrom, B. Giros, M. Leboyer, C. Gillberg and T. Bourgeron (2003). "Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism." *Nat Genet* **34**: 27-9.

James, R., R. R. Adams, S. Christie, S. R. Buchanan, D. J. Porteous and J. K. Millar (2004). "Disrupted in Schizophrenia 1 (DISC1) is a multicompartmentalized protein that predominantly localizes to mitochondria." *Mol Cell Neurosci* **26**: 112-22.

Just, M. A., V. L. Cherkassky, T. A. Keller and N. J. Minshew (2004). "Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity." *Brain* **127**: 1811-21.

Kadesjo, B., C. Gillberg and B. Hagberg (1999). "Brief report: autism and Asperger syndrome in seven-year-old children: a total population study." *J Autism Dev Disord* **29**: 327-31.

Kallio, S. P., E. Jakkula, S. Purcell, M. Suvela, K. Koivisto, P. J. Tienari, I. Elovaara, T. Pirttila, M. Reunanen, D. Bronnikov, M. Viander, S. Meri, J. Hillert, F. Lundmark, H. F. Harbo, A. R. Lorentzen, P. L. De Jager, M. J. Daly, D. A. Hafler, A. Palotie, L. Peltonen and J. Saarela (2009). "Use of a genetic isolate to identify rare disease variants: C7 on 5p associated with MS." *Hum Mol Genet* **18**: 1670-83.

Kamiya, A., K. Kubo, T. Tomoda, M. Takaki, R. Youn, Y. Ozeki, N. Sawamura, U. Park, C. Kudo, M. Okawa, C. A. Ross, M. E. Hatten, K. Nakajima and A. Sawa (2005). "A schizophrenia-associated mutation of DISC1 perturbs cerebral cortex development." *Nat Cell Biol* **7**: 1167-78.

Kamiya, A., T. Tomoda, J. Chang, M. Takaki, C. Zhan, M. Morita, M. B. Cascio, S. Elashvili, H. Koizumi, Y. Takanezawa, F. Dickerson, R. Yolken, H. Arai and A. Sawa (2006). "DISC1-NDEL1/NUDEL protein interaction, an essential component for neurite outgrowth, is modulated by genetic variations of DISC1." *Hum Mol Genet* **15**: 3313-23.

Kanner, L. (1943). "Autistic disturbances of affective contact." *Nervous Child* **2**: 217-250.

Kanner, L. (1949). "Problems of nosology and psychodynamics of early infantile autism." *Am J Orthopsychiatry* **19**: 416-26.

Kestilä, M., U. Lenkkeri, M. Mannikko, J. Lamerdin, P. McCready, H. Putaala, V. Ruotsalainen, T. Morita, M. Nissinen, R. Herva, C. E. Kashtan, L. Peltonen, C. Holmberg, A. Olsen and K. Tryggvason (1998). "Positionally cloned gene for a novel glomerular protein--nephrin--is mutated in congenital nephrotic syndrome." *Mol Cell* **1**: 575-82.

Kilpinen, H., T. Ylisaukko-oja, K. Rehnstrom, E. Gaal, J. A. Turunen, E. Kempas, L. von Wendt, T. Varilo and L. Peltonen (2009). "Linkage and linkage disequilibrium scan for autism loci in an extended pedigree from Finland." *Hum Mol Genet* **18**: 2912-21.

Kim, A. H., M. Reimers, B. Maher, V. Williamson, O. McMichael, J. L. McClay, E. J. van den Oord, B. P. Riley, K. S. Kendler and V. I. Vladimirov (2010). "MicroRNA expression profiling in the prefrontal cortex of individuals affected with schizophrenia and bipolar disorders." *Schizophr Res*.

Kim, V. N., J. Han and M. C. Siomi (2009). "Biogenesis of small RNAs in animals." *Nat Rev Mol Cell Biol* **10**: 126-39.

Knickmeyer, R. C. and S. Baron-Cohen (2006). "Fetal testosterone and sex differences in typical social development and in autism." *J Child Neurol* **21**: 825-45.

Koike, H., P. A. Arguello, M. Kvajo, M. Karayiorgou and J. A. Gogos (2006). "Disc1 is mutated in the 129S6/SvEv strain and modulates working memory in mice." *Proc Natl Acad Sci U S A* **103**: 3693-7.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S. T. Palsson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher and K. Stefansson (2002). "A high-resolution recombination map of the human genome." *Nat Genet* **31**: 241-7.

Koshino, H., R. K. Kana, T. A. Keller, V. L. Cherkassky, N. J. Minshew and M. A. Just (2008). "fMRI investigation of working memory for faces in autism: visual coding and underconnectivity with frontal areas." *Cereb Cortex* **18**: 289-300.

Krey, J. F. and R. E. Dolmetsch (2007). "Molecular mechanisms of autism: a possible role for Ca2+ signaling." *Curr Opin Neurobiol* **17**: 112-9.

Kristiansson, K., J. Naukkarinen and L. Peltonen (2008). "Isolated populations and complex disease gene identification." *Genome Biol* **9**: 109.

Krol, J., I. Loedige and W. Filipowicz (2010). "The widespread regulation of microRNA biogenesis, function and decay." *Nat Rev Genet* **11**: 597-610.

Kumar, R. A., S. KaraMohamed, J. Sudi, D. F. Conrad, C. Brune, J. A. Badner, T. C. Gilliam, N. J. Nowak, E. H. Cook, Jr., W. B. Dobyns and S. L. Christian (2008). "Recurrent 16p11.2 microdeletions in autism." *Hum Mol Genet* **17**: 628-38.

Kumar, S. and S. Subramanian (2002). "Mutation rates in mammalian genomes." *Proc Natl Acad Sci U S A* **99**: 803-8.

Kuo, T. Y., C. J. Hong, H. L. Chien and Y. P. Hsueh (2010). "X-linked mental retardation gene CASK interacts with Bcl11A/CTIP1 and regulates axon branching and outgrowth." *J Neurosci Res* **88**: 2364-73.

Kuo, T. Y., C. J. Hong and Y. P. Hsueh (2009). "Bcl11A/CTIP1 regulates expression of DCC and MAP1b in control of axon branching and dendrite outgrowth." *Mol Cell Neurosci* **42**: 195-207.

Kwan, T., D. Benovoy, C. Dias, S. Gurd, C. Provencher, P. Beaulieu, T. J. Hudson, R. Sladek and J. Majewski (2008). "Genome-wide analysis of transcript isoform variation in humans." *Nat Genet* **40**: 225-31.

Laitinen, T., A. Polvi, P. Rydman, J. Vendelin, V. Pulkkinen, P. Salmikangas, S. Makela, M. Rehn, A. Pirskanen, A. Rautanen, M. Zucchelli, H. Gullsten, M. Leino, H. Alenius, T. Petays, T. Haahtela, A. Laitinen, C. Laprise, T. J. Hudson, L. A. Laitinen and J. Kere (2004). "Characterization of a common susceptibility locus for asthma-related traits." *Science* **304**: 300-4.

Lamb, J. A., G. Barnby, E. Bonora, N. Sykes, E. Bacchelli, F. Blasi, E. Maestrini, J. Broxholme, J. Tzenova, D. Weeks, A. J. Bailey and A. P. Monaco (2005). "Analysis of IMGSAC autism susceptibility loci: evidence for sex limited and parent of origin specific effects." *J Med Genet* **42**: 132-7.

Lampi, K. M., A. Sourander, M. Gissler, S. Niemela, K. Rehnstrom, E. Pulkkinen, L. Peltonen and L. Von Wendt (2010). "Brief Report: Validity of Finnish Registry-Based Diagnoses of Autism with the ADI-R." *Acta Paediatr*.

Lander, E. S. (1996). "The new genomics: global views of biology." *Science* **274**: 536-9.

Landrigan, P. J. (2010). "What causes autism? Exploring the environmental contribution." *Curr Opin Pediatr* **22**: 219-25.

Lange, E. M. and K. Lange (2004). "Powerful allele sharing statistics for nonparametric linkage analysis." *Hum Hered* **57**: 49-58.

Laumonnier, F., F. Bonnet-Brilhault, M. Gomot, R. Blanc, A. David, M. P. Moizard, M. Raynaud, N. Ronce, E. Lemonnier, P. Calvas, B. Laudier, J. Chelly, J. P. Fryns, H. H. Ropers, B. C. Hamel, C. Andres, C. Barthelemy, C. Moraine and S. Briault (2004). "X-Linked Mental Retardation and Autism Are Associated with a Mutation in the NLGN4 Gene, a Member of the Neuroligin Family." *Am J Hum Genet* **74**: 552-7.

Leid, M., J. E. Ishmael, D. Avram, D. Shepherd, V. Fraulob and P. Dolle (2004). "CTIP1 and CTIP2 are differentially expressed during mouse embryogenesis." *Gene Expr Patterns* **4**: 733-9.

Lennon, G., C. Auffray, M. Polymeropoulos and M. B. Soares (1996). "The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression." *Genomics* **33**: 151-2.

Lenzen, K. P., A. Heils, S. Lorenz, A. Hempelmann, S. Hofels, F. W. Lohoff, B. Schmitz and T. Sander (2005). "Supportive evidence for an allelic association of the human KCNJ10 potassium channel gene with idiopathic generalized epilepsy." *Epilepsy Res* **63**: 113-8.

Lesnick, T. G., S. Papapetropoulos, D. C. Mash, J. Ffrench-Mullen, L. Shehadeh, M. de Andrade, J. R. Henley, W. A. Rocca, J. E. Ahlskog and D. M. Maraganore (2007). "A genomic pathway approach to a complex disease: axon guidance and Parkinson disease." *PLoS Genet* **3**: e98.

Levinson, G. and G. A. Gutman (1987). "High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12." *Nucleic Acids Res* **15**: 5323-38.

Lewis, B. P., C. B. Burge and D. P. Bartel (2005). "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." *Cell* **120**: 15-20.

Liu, J., D. R. Nyholt, P. Magnussen, E. Parano, P. Pavone, D. Geschwind, C. Lord, P. Iversen, J. Hoh, J. Ott and T. C. Gilliam (2001). "A genomewide screen for autism susceptibility loci." *Am J Hum Genet* **69**: 327-40.

Liu, Y. L., C. S. Fann, C. M. Liu, W. J. Chen, J. Y. Wu, S. I. Hung, C. H. Chen, Y. S. Jou, S. K. Liu, T. J. Hwang, M. H. Hsieh, W. C. Ouyang, H. Y. Chan, J. J. Chen, W. C. Yang, C. Y. Lin, S. F. Lee and H. G. Hwu (2006). "A single nucleotide polymorphism fine mapping study of chromosome 1q42.1 reveals the vulnerability genes for schizophrenia, GNPAT and DISC1: Association with impairment of sustained attention." *Biol Psychiatry* **60**: 554-62.

Lord, C., M. Rutter, S. Goode, J. Heemsbergen, H. Jordan, L. Mawhood and E. Schopler (1989). "Autism diagnostic observation schedule: a standardized observation of communicative and social behavior." *J Autism Dev Disord* **19**: 185-212.

Lord, C., M. Rutter and A. Le Couteur (1994). "Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders." *J Autism Dev Disord* **24**: 659-85.

Lu, J., G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz and T. R. Golub (2005). "MicroRNA expression profiles classify human cancers." *Nature* **435**: 834-8.

Ma, D., D. Salyakina, J. M. Jaworski, I. Konidari, P. L. Whitehead, A. N. Andersen, J. D. Hoffman, S. H. Slifer, D. J. Hedges, H. N. Cukier, A. J. Griswold, J. L. McCauley, G. W. Beecham, H. H. Wright, R. K. Abramson, E. R. Martin, J. P. Hussman, J. R. Gilbert, M. L. Cuccaro, J. L. Haines and M. A. Pericak-Vance (2009). "A genome-wide association study of autism reveals a common novel risk locus at 5p14.1." *Ann Hum Genet* **73**: 263-73.

Mackie, S., J. K. Millar and D. J. Porteous (2007). "Role of DISC1 in neural development and schizophrenia." *Curr Opin Neurobiol* **17**: 95-102.

Maeda, K., E. Nwulia, J. Chang, R. Balkissoon, K. Ishizuka, H. Chen, P. Zandi, M. G. McInnis and A. Sawa (2006). "Differential expression of disrupted-in-schizophrenia (DISC1) in bipolar disorder." *Biol Psychiatry* **60**: 929-35.

Mao, Y., X. Ge, C. L. Frank, J. M. Madison, A. N. Koehler, M. K. Doud, C. Tassa, E. M. Berry, T. Soda, K. K. Singh, T. Biechele, T. L. Petryshen, R. T. Moon, S. J. Haggarty and L. H. Tsai (2009). "Disrupted in schizophrenia 1 regulates neuronal progenitor proliferation via modulation of GSK3beta/beta-catenin signaling." *Cell* **136**: 1017-31.

Marshall, C. R., A. Noor, J. B. Vincent, A. C. Lionel, L. Feuk, J. Skaug, M. Shago, R. Moessner, D. Pinto, Y. Ren, B. Thiruvahindrapduram, A. Fiebig, S. Schreiber, J. Friedman, C. E. Ketelaars, Y. J. Vos, C. Ficicioglu, S. Kirkpatrick, R. Nicolson, L. Sloman, A. Summers, C. A. Gibbons, A. Teebi, D. Chitayat, R. Weksberg, A. Thompson, C. Vardy, V. Crosbie, S. Luscombe, R. Baatjes, L. Zwaigenbaum, W. Roberts, B. Fernandez, P. Szatmari and S. W. Scherer (2008). "Structural variation of chromosomes in autism spectrum disorder." *Am J Hum Genet* **82**: 477-88.

Mattila, M. L., M. Kielinen, K. Jussila, S. L. Linna, R. Bloigu, H. Ebeling and I. Moilanen (2007). "An epidemiological and diagnostic study of Asperger syndrome according to four sets of diagnostic criteria." *J Am Acad Child Adolesc Psychiatry* **46**: 636-46.

McCauley, J. L., C. Li, L. Jiang, L. M. Olson, G. Crockett, K. Gainer, S. E. Folstein, J. L. Haines and J. S. Sutcliffe (2005). "Genome-wide and Ordered-Subset linkage analyses provide support for autism loci on 17q and 19p with evidence of phenotypic and interlocus genetic correlates." *BMC Med Genet* **6**: 1.

Melin, M., B. Carlsson, H. Anckarsater, M. Rastam, C. Betancur, A. Isaksson, C. Gillberg and N. Dahl (2006). "Constitutional downregulation of SEMA5A expression in autism." *Neuropsychobiology* **54**: 64-9.

Menashe, I., D. Maeder, M. Garcia-Closas, J. D. Figueroa, S. Bhattacharjee, M. Rotunno, P. Kraft, D. J. Hunter, S. J. Chanock, P. S. Rosenberg and N. Chatterjee "Pathway Analysis of Breast Cancer Genome-Wide Association Study Highlights Three Pathways and One Canonical Signaling Cascade." *Cancer Res.*

Menzel, S., C. Garner, I. Gut, F. Matsuda, M. Yamaguchi, S. Heath, M. Foglio, D. Zelenika, A. Boland, H. Rooks, S. Best, T. D. Spector, M. Farrall, M. Lathrop and S. L. Thein (2007). "A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15." *Nat Genet* **39**: 1197-9.

Meyer, G., F. Varoqueaux, A. Neeb, M. Oschlies and N. Brose (2004). "The complexity of PDZ domain-mediated interactions at glutamatergic synapses: a case study on neuroligin." *Neuropharmacology* **47**: 724-33.

Millar, J. K., S. Christie, S. Anderson, D. Lawson, D. Hsiao-Wei Loh, R. S. Devon, B. Arveiler, W. J. Muir, D. H. Blackwood and D. J. Porteous (2001). "Genomic structure and localisation within a linkage hotspot of Disrupted In Schizophrenia 1, a gene disrupted by a translocation segregating with schizophrenia." *Mol Psychiatry* **6**: 173-8.

Millar, J. K., S. Christie and D. J. Porteous (2003). "Yeast two-hybrid screens implicate DISC1 in brain development and function." *Biochem Biophys Res Commun* **311**: 1019-25.

Millar, J. K., S. Christie, C. A. Semple and D. J. Porteous (2000a). "Chromosomal location and genomic structure of the human translin-associated factor X gene (TRAX; TSNAX) revealed by intergenic splicing to DISC1, a gene disrupted by a translocation segregating with schizophrenia." *Genomics* **67**: 69-77.

Millar, J. K., B. S. Pickard, S. Mackie, R. James, S. Christie, S. R. Buchanan, M. P. Malloy, J. E. Chubb, E. Huston, G. S. Baillie, P. A. Thomson, E. V. Hill, N. J. Brandon, J. C. Rain, L. M. Camargo, P. J. Whiting, M. D. Houslay, D. H. Blackwood, W. J. Muir and D. J. Porteous (2005). "DISC1 and PDE4B are interacting genetic factors in schizophrenia that regulate cAMP signaling." *Science* **310**: 1187-91.

Millar, J. K., J. C. Wilson-Annan, S. Anderson, S. Christie, M. S. Taylor, C. A. Semple, R. S. Devon, D. M. Clair, W. J. Muir, D. H. Blackwood and D. J. Porteous (2000b). "Disruption of two novel genes by a translocation co-segregating with schizophrenia." *Hum Mol Genet* **9**: 1415-23.

Min, J. L., A. Barrett, T. Watts, F. H. Pettersson, H. E. Lockstone, C. M. Lindgren, J. M. Taylor, M. Allen, K. T. Zondervan and M. I. McCarthy (2010). "Variability of gene expression profiles in human blood and lymphoblastoid cell lines." *BMC Genomics* **11**: 96.

Miyasaka, H., B. K. Choudhury, E. W. Hou and S. S. Li (1993). "Molecular cloning and expression of mouse and human cDNA encoding AES and ESG proteins with strong similarity to Drosophila enhancer of split groucho protein." *Eur J Biochem* **216**: 343-52.

Miyoshi, K., A. Honda, K. Baba, M. Taniguchi, K. Oono, T. Fujita, S. Kuroda, T. Katayama and M. Tohyama (2003). "Disrupted-In-Schizophrenia 1, a candidate gene for schizophrenia, participates in neurite outgrowth." *Mol Psychiatry* **8**: 685-94.

Moessner, R., C. R. Marshall, J. S. Sutcliffe, J. Skaug, D. Pinto, J. Vincent, L. Zwaigenbaum, B. Fernandez, W. Roberts, P. Szatmari and S. W. Scherer (2007). "Contribution of SHANK3 Mutations to Autism Spectrum Disorder." *Am J Hum Genet* **81**: 1289-97.

Moffatt, M. F., M. Kabesch, L. Liang, A. L. Dixon, D. Strachan, S. Heath, M. Depner, A. von Berg, A. Bufe, E. Rietschel, A. Heinzmann, B. Simma, T. Frischer, S. A. Willis-Owen, K. C. Wong, T. Illig, C. Vogelberg, S. K. Weiland, E. von Mutius, G. R. Abecasis, M. Farrall, I. G. Gut, G. M. Lathrop and W. O. Cookson (2007). "Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma." *Nature* **448**: 470-3.

Mootha, V. K., C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler and L. C. Groop (2003). "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes." *Nat Genet* **34**: 267-73.

Morris, J. A., G. Kandpal, L. Ma and C. P. Austin (2003). "DISC1 (Disrupted-In-Schizophrenia 1) is a centrosome-associated protein that interacts with MAP1A, MIPT3, ATF4/5 and NUDEL: regulation and loss of interaction with mutation." *Hum Mol Genet* **12**: 1591-608.

Morrow, E. M., S. Y. Yoo, S. W. Flavell, T. K. Kim, Y. Lin, R. S. Hill, N. M. Mukaddes, S. Balkhy, G. Gascon, A. Hashmi, S. Al-Saad, J. Ware, R. M. Joseph, R. Greenblatt, D. Gleason, J. A. Ertelt, K. A. Apse, A. Bodell, J. N. Partlow, B. Barry, H. Yao, K. Markianos, R. J. Ferland, M. E. Greenberg and C. A. Walsh (2008). "Identifying autism loci and genes by tracing recent shared ancestry." *Science* **321**: 218-23.

Morton, N. E. (1955). "Sequental tests for the detection of linkage." *Am J Hum Genet* **7**: 277-318.

Najm, J., D. Horn, I. Wimplinger, J. A. Golden, V. V. Chizhikov, J. Sudi, S. L. Christian, R. Ullmann, A. Kuechler, C. A. Haas, A. Flubacher, L. R. Charnas, G. Uyanik, U. Frank, E. Klopocki, W. B. Dobyns and K. Kutsche (2008). "Mutations of CASK cause an X-linked brain malformation phenotype with microcephaly and hypoplasia of the brainstem and cerebellum." *Nat Genet.*

Nakata, K., B. K. Lipska, T. M. Hyde, T. Ye, E. N. Newburn, Y. Morita, R. Vakkalanka, M. Barenboim, Y. Sei, D. R. Weinberger and J. E. Kleinman (2009). "DISC1 splice variants are upregulated in schizophrenia and associated with risk polymorphisms." *Proc Natl Acad Sci U S A* **106**: 15873-8.

Ng, S. B., K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure and M. J. Bamshad (2010). "Exome sequencing identifies the cause of a mendelian disorder." *Nat Genet* **42**: 30-5.

Nica, A. C., S. B. Montgomery, A. S. Dimas, B. E. Stranger, C. Beazley, I. Barroso and E. T. Dermitzakis (2010). "Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations." *PLoS Genet* **6**: e1000895.

Nieminen-Von Wendt, T. (2004). On the origin and diagnosis of Asperger syndrome: A clinical, neuroimaging and genetic study. Helsinki, University of Helsinki.

Nishimura, Y., C. L. Martin, A. Vazquez-Lopez, S. J. Spence, A. I. Alvarez-Retuerto, M. Sigman, C. Steindler, S. Pellegrini, N. C. Schanen, S. T. Warren and D. H. Geschwind (2007).

"Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways." *Hum Mol Genet* **16**: 1682-98.

Nomura, T., M. Kimura, T. Horii, S. Morita, H. Soejima, S. Kudo and I. Hatada (2008). "MeCP2-dependent repression of an imprinted miR-184 released by depolarization." *Hum Mol Genet* **17**: 1192-9.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens and C. D. Bustamante (2008). "Genes mirror geography within Europe." *Nature* **456**: 98-101.

Numata, S., J. Iga, M. Nakataki, S. Tayoshi, K. Taniguchi, S. Sumitani, M. Tomotake, T. Tanahashi, M. Itakura, Y. Kamegaya, M. Tatsumi, A. Sano, T. Asada, H. Kunugi, S. Ueno and T. Ohmori (2009). "Gene expression and association analyses of the phosphodiesterase 4B (PDE4B) gene in major depressive disorder in the Japanese population." *Am J Med Genet B Neuropsychiatr Genet* **150B**: 527-34.

Nyholt, D. R., K. S. LaForge, M. Kallela, K. Alakurtti, V. Anttila, M. Farkkila, E. Hamalainen, J. Kaprio, M. A. Kaunisto, A. C. Heath, G. W. Montgomery, H. Gobel, U. Todt, M. D. Ferrari, L. J. Launer, R. R. Frants, G. M. Terwindt, B. de Vries, W. M. Verschuren, J. Brand, T. Freilinger, V. Pfaffenrath, A. Straube, D. G. Ballinger, Y. Zhan, M. J. Daly, D. R. Cox, M. Dichgans, A. M. van den Maagdenberg, C. Kubisch, N. G. Martin, M. Wessman, L. Peltonen and A. Palotie (2008). "A high-density association screen of 155 ion transport genes for involvement with common migraine." *Hum Mol Genet* **17**: 3318-31.

O'Connell, J. R. and D. E. Weeks (1998). "PedCheck: a program for identification of genotype incompatibilities in linkage analysis." *Am J Hum Genet* **63**: 259-66.

O'Dushlaine, C., E. Kenny, E. Heron, G. Donohoe, M. Gill, D. Morris and A. Corvin (2010). "Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility." *Mol Psychiatry Feb 16 [Epub ahead of print]*.

O'Dushlaine, C., E. Kenny, E. A. Heron, R. Segurado, M. Gill, D. W. Morris and A. Corvin (2009). "The SNP ratio test: pathway analysis of genome-wide association datasets." *Bioinformatics* **25**: 2762-3.

Ozeki, Y., T. Tomoda, J. Kleiderlein, A. Kamiya, L. Bord, K. Fujii, M. Okawa, N. Yamada, M. E. Hatten, S. H. Snyder, C. A. Ross and A. Sawa (2003). "Disrupted-in-Schizophrenia-1 (DISC-1): mutant truncation prevents binding to NudE-like (NUDEL) and inhibits neurite outgrowth." *Proc Natl Acad Sci U S A* **100**: 289-94.

Pagnamenta, A. T., E. Bacchelli, M. V. de Jonge, G. Mirza, T. S. Scerri, F. Minopoli, A. Chiocchetti, K. U. Ludwig, P. Hoffmann, S. Paracchini, E. Lowy, D. H. Harold, J. A. Chapman, S. M. Klauck, F. Poustka, R. H. Houben, W. G. Staal, R. A. Ophoff, M. C. O'Donovan, J. Williams, M. M. Nothen, G. Schulte-Korne, P. Deloukas, J. Ragoussis, A. J. Bailey, E. Maestrini and A. P. Monaco (2010). "Characterization of a Family with Rare Deletions in CNTNAP5 and DOCK4 Suggests Novel Risk Loci for Autism and Dyslexia." *Biol Psychiatry*.

Pajukanta, P., H. E. Lilja, J. S. Sinsheimer, R. M. Cantor, A. J. Lusis, M. Gentile, X. J. Duan, A. Soro-Paavonen, J. Naukkarinen, J. Saarela, M. Laakso, C. Ehnholm, M. R. Taskinen and L. Peltonen (2004). "Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1)." *Nat Genet* **36**: 371-6.

Palo, O. M., M. Antila, K. Silander, W. Hennah, H. Kilpinen, P. Soronen, A. Tuulio-Henriksson, T. Kieseppa, T. Partonen, J. Lonnqvist, L. Peltonen and T. Paunio (2007). "Association of distinct allelic haplotypes of DISC1 with psychotic and bipolar spectrum disorders and with underlying cognitive impairments." *Hum Mol Genet* **16**: 3517-28.

Peltonen, L., A. Jalanko and T. Varilo (1999). "Molecular genetics of the Finnish disease heritage." *Hum Mol Genet* **8**: 1913-23.

Peltonen, L., A. Palotie and K. Lange (2000). "Use of population isolates for mapping complex traits." *Nat Rev Genet* **1**: 182-90.

Persico, A. M., L. D'Agruma, N. Maiorano, A. Totaro, R. Militerni, C. Bravaccio, T. H. Wassink, C. Schneider, R. Melmed, S. Trillo, F. Montecchi, M. Palermo, T. Pascucci, S. Puglisi-Allegra, K. L. Reichelt, M. Conciatori, R. Marino, C. C. Quattrocchi, A. Baldi, L. Zelante, P. Gasparini and F. Keller (2001). "Reelin gene alleles and haplotypes as a factor predisposing to autistic disorder." *Mol Psychiatry* **6**: 150-9.

Phillips, M. S., R. Lawrence, R. Sachidanandam, A. P. Morris, D. J. Balding, M. A. Donaldson, J. F. Studebaker, W. M. Ankener, S. V. Alfisi, F. S. Kuo, A. L. Camisa, V. Pazorov, K. E. Scott, B. J. Carey, J. Faith, G. Katari, H. A. Bhatti, J. M. Cyr, V. Derohannessian, C. Elosua, A. M. Forman, N. M. Grecco, C. R. Hock, J. M. Kuebler, J. A. Lathrop, M. A. Mockler, E. P. Nachtman, S. L. Restine, S. A. Varde, M. J. Hozza, C. A. Gelfand, J. Broxholme, G. R. Abecasis, M. T. Boyce-Jacino and L. R. Cardon (2003). "Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots." *Nat Genet* **33**: 382-7.

Pickles, A., P. Bolton, H. Macdonald, A. Bailey, A. Le Couteur, C. H. Sim and M. Rutter (1995). "Latent-class analysis of recurrence risks for complex phenotypes with selection and measurement error: a twin and family history study of autism." *Am J Hum Genet* **57**: 717-26.

Pietilainen, K. H., J. Naukkarinen, A. Rissanen, J. Saharinen, P. Ellonen, H. Keranen, A. Suomalainen, A. Gotz, T. Suortti, H. Yki-Jarvinen, M. Oresic, J. Kaprio and L. Peltonen (2008). "Global transcript profiles of fat in monozygotic twins discordant for BMI: pathways behind acquired obesity." *PLoS Med* **5**: e51.

Pinto, D., A. T. Pagnamenta, L. Klei, R. Anney, D. Merico, R. Regan, J. Conroy, T. R. Magalhaes, C. Correia, B. S. Abrahams, J. Almeida, E. Bacchelli, G. D. Bader, A. J. Bailey, G. Baird, A. Battaglia, T. Berney, N. Bolshakova, S. Bolte, P. F. Bolton, T. Bourgeron, S. Brennan, J. Brian, S. E. Bryson, A. R. Carson, G. Casallo, J. Casey, B. H. Chung, L. Cochrane, C. Corsello, E. L. Crawford, A. Crossett, C. Cytrynbaum, G. Dawson, M. de Jonge, R. Delorme, I. Drmic, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, E. Fombonne, C. M. Freitag, J. Gilbert, C. Gillberg, J. T. Glessner, J. Goldberg, A. Green, J. Green, S. J. Guter, H. Hakonarson, E. A. Heron, M. Hill, R. Holt, J. L. Howe, G. Hughes, V. Hus, R. Igliozzi, C. Kim, S. M. Klauck, A. Kolevzon, O. Korvatska, V. Kustanovich, C. M. Lajonchere, J. A. Lamb, M. Laskawiec, M. Leboyer, A. Le Couteur, B. L. Leventhal, A. C. Lionel, X. Q. Liu, C. Lord, L. Lotspeich, S. C. Lund, E. Maestrini, W. Mahoney, C. Mantoulan, C. R. Marshall, H. McConachie, C. J. McDougle, J. McGrath, W. M. McMahon, A. Merikangas, O. Migita, N. J. Minshew, G. K. Mirza, J. Munson, S. F. Nelson, C. Noakes, A. Noor, G. Nygren, G. Oliveira, K. Papanikolaou, J. R. Parr, B. Parrini, T. Paton, A. Pickles and M. Pilorge (2010). "Functional impact of global rare copy number variation in autism spectrum disorders." *Nature* **466**: 368-72.

Plagnol, V., E. Uz, C. Wallace, H. Stevens, D. Clayton, T. Ozcelik and J. A. Todd (2008). "Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses." *PLoS One* **3**: e2966.

Poliak, S., L. Gollan, R. Martinez, A. Custer, S. Einheber, J. L. Salzer, J. S. Trimmer, P. Shrager and E. Peles (1999). "Caspr2, a new member of the neurexin superfamily, is localized at the juxtaparanodes of myelinated axons and associates with K+ channels." *Neuron* **24**: 1037-47.

Purcell, A. E., O. H. Jeon, A. W. Zimmerman, M. E. Blue and J. Pevsner (2001). "Postmortem brain abnormalities of the glutamate neurotransmitter system in autism." *Neurology* **57**: 1618-28.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly and P. C. Sham (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." *Am J Hum Genet* **81**: 559-75.

Reddy, K. S. (2005). "Cytogenetic abnormalities and fragile-X syndrome in Autism Spectrum Disorder." *BMC Med Genet* **6**: 3.

Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer and M. E. Hurles (2006). "Global variation in copy number in the human genome." *Nature* **444**: 444-54.

Rehnström, K., T. Ylisaukko-oja, T. Nieminen-von Wendt, S. Sarenius, T. Källman, E. Kempas, L. von Wendt, L. Peltonen and I. Järvelä (2006). "Independent replication and initial fine mapping of 3p21-24 in Asperger syndrome." *J Med Genet* **43**: e6.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward and E. S. Lander (2001). "Linkage disequilibrium in the human genome." *Nature* **411**: 199-204.

Reich, D. E. and E. S. Lander (2001). "On the allelic spectrum of human disease." *Trends Genet* **17**: 502-10.

Reichenberg, A., R. Gross, M. Weiser, M. Bresnahan, J. Silverman, S. Harlap, J. Rabinowitz, C. Shulman, D. Malaspina, G. Lubin, H. Y. Knobler, M. Davidson and E. Susser (2006). "Advancing paternal age and autism." *Arch Gen Psychiatry* **63**: 1026-32.

Risch, N., D. Spiker, L. Lotspeich, N. Nouri, D. Hinds, J. Hallmayer, L. Kalaydjieva, P. McCague, S. Dimiceli, T. Pitts, L. Nguyen, J. Yang, C. Harper, D. Thorpe, S. Vermeer, H. Young, J. Hebert, A. Lin, J. Ferguson, C. Chiotti, S. Wiese-Slater, T. Rogers, B. Salmon, P. Nicholas, R. M. Myers and et al. (1999). "A genomic screen of autism: evidence for a multilocus etiology." *Am J Hum Genet* **65**: 493-507.

Ritvo, E. R., A. Mason-Brothers, B. J. Freeman, C. Pingree, W. R. Jenson, W. M. McMahon, P. B. Petersen, L. B. Jorde, A. Mo and A. Ritvo (1990). "The UCLA-University of Utah epidemiologic survey of autism: the etiologic role of rare diseases." *Am J Psychiatry* **147**: 1614-21.

Ross, C. A., R. L. Margolis, S. A. Reading, M. Pletnikov and J. T. Coyle (2006). "Neurobiology of schizophrenia." *Neuron* **52**: 139-53.

Roth, M., B. Bonev, J. Lindsay, R. Lea, N. Panagiotaki, C. Houart and N. Papalopulu (2010). "FoxG1 and TLE2 act cooperatively to regulate ventral telencephalon formation." *Development* **137**: 1553-62.

Rubenstein, J. L. and M. M. Merzenich (2003). "Model of autism: increased ratio of excitation/inhibition in key neural systems." *Genes Brain Behav* **2**: 255-67.

Rutter, M. (1968). "Concepts of autism: a review of research." *J Child Psychol Psychiatry* **9**: 1-25.

Sachs, N. A., A. Sawa, S. E. Holmes, C. A. Ross, L. E. DeLisi and R. L. Margolis (2005). "A frameshift mutation in Disrupted in Schizophrenia 1 in an American family with schizophrenia and schizoaffective disorder." *Mol Psychiatry* **10**: 758-64.

Sajantila, A., A. H. Salem, P. Savolainen, K. Bauer, C. Gierig and S. Paabo (1996). "Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population." *Proc Natl Acad Sci U S A* **93**: 12035-9.

Salmela, E., T. Lappalainen, I. Fransson, P. M. Andersen, K. Dahlman-Wright, A. Fiebig, P. Sistonen, M. L. Savontaus, S. Schreiber, J. Kere and P. Lahermo (2008). "Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe." *PLoS One* **3**: e3519.

Sanger, F. and A. R. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase." *J Mol Biol* **94**: 441-8.

Sankaran, V. G., T. F. Menne, J. Xu, T. E. Akie, G. Lettre, B. Van Handel, H. K. Mikkola, J. N. Hirschhorn, A. B. Cantor and S. H. Orkin (2008). "Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A." *Science* **322**: 1839-42.

Sarachana, T., R. Zhou, G. Chen, H. K. Manji and V. W. Hu (2010). "Investigation of post-transcriptional gene regulatory networks associated with autism spectrum disorders by microRNA expression profiling of lymphoblastoid cell lines." *Genome Med* **2**: 23.

Schellenberg, G. D., G. Dawson, Y. J. Sung, A. Estes, J. Munson, E. Rosenthal, J. Rothstein, P. Flodman, M. Smith, H. Coon, L. Leong, C. E. Yu, C. Stodgell, P. M. Rodier, M. A. Spence, N. Minshew, W. M. McMahon and E. M. Wijsman (2006). "Evidence for multiple loci from a genome scan of autism kindreds." *Mol Psychiatry* **11**: 1049-60, 979.

Schoeb, D. S., G. Chernin, S. F. Heeringa, V. Matejas, S. Held, V. Vega-Warner, D. Bockenhauer, C. N. Vlangos, K. N. Moorani, T. J. Neuhaus, J. A. Kari, J. Macdonald, P. Saisawat, S. Ashraf, B. Ovunc, M. Zenker and F. Hildebrandt (2010). "Nineteen novel NPHS1 mutations in a worldwide cohort of patients with congenital nephrotic syndrome (CNS)." *Nephrol Dial Transplant* **25**: 2970-6.

Schumann, C. M., J. Hamstra, B. L. Goodlin-Jones, L. J. Lotspeich, H. Kwon, M. H. Buonocore, C. R. Lammers, A. L. Reiss and D. G. Amaral (2004). "The amygdala is enlarged in children but not adolescents with autism; the hippocampus is enlarged at all ages." *J Neurosci* **24**: 6392-401.

Schurov, I. L., E. J. Handford, N. J. Brandon and P. J. Whiting (2004). "Expression of disrupted in schizophrenia 1 (DISC1) protein in the adult and developing mouse brain indicates its role in neurodevelopment." *Mol Psychiatry* **9**: 1100-10.

Sebat, J., B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y. H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimaki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M. C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye and M. Wigler (2007). "Strong association of de novo copy number mutations with autism." *Science* **316**: 445-9.

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg and M. Wigler (2004). "Large-scale copy number polymorphism in the human genome." *Science* **305**: 525-8.

Serajee, F. J., H. Zhong and A. H. Mahbubul Huq (2006). "Association of Reelin gene polymorphisms with autism." *Genomics* **87**: 75-83.

Service, S., J. DeYoung, M. Karayiorgou, J. L. Roos, H. Pretorious, G. Bedoya, J. Ospina, A. Ruiz-Linares, A. Macedo, J. A. Palha, P. Heutink, Y. Aulchenko, B. Oostra, C. van Duijn, M. R. Jarvelin, T. Varilo, L. Peddle, P. Rahman, G. Piras, M. Monne, S. Murray, L. Galver, L. Peltonen, C. Sabatti, A. Collins and N. Freimer (2006). "Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies." *Nat Genet* **38**: 556-60.

Sethupathy, P., C. Borel, M. Gagnebin, G. R. Grant, S. Deutsch, T. S. Elton, A. G. Hatzigeorgiou and S. E. Antonarakis (2007). "Human microRNA-155 on chromosome 21 differentially interacts with its polymorphic target in the AGTR1 3' untranslated region: a mechanism

for functional single-nucleotide polymorphisms related to phenotypes." *Am J Hum Genet* **81**: 405-13.

Shao, Y., C. M. Wolpert, K. L. Raiford, M. M. Menold, S. L. Donnelly, S. A. Ravan, M. P. Bass, C. McClain, L. von Wendt, J. M. Vance, R. H. Abramson, H. H. Wright, A. Ashley-Koch, J. R. Gilbert, R. G. DeLong, M. L. Cuccaro and M. A. Pericak-Vance (2002). "Genomic screen and follow-up analysis for autistic disorder." *Am J Med Genet* **114**: 99-105.

Skaar, D. A., Y. Shao, J. L. Haines, J. E. Stenger, J. Jaworski, E. R. Martin, G. R. DeLong, J. H. Moore, J. L. McCauley, J. S. Sutcliffe, A. E. Ashley-Koch, M. L. Cuccaro, S. E. Folstein, J. R. Gilbert and M. A. Pericak-Vance (2005). "Analysis of the RELN gene as a genetic risk factor for autism." *Mol Psychiatry* **10**: 563-71.

Smyth, G. K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Stat Appl Genet Mol Biol* **3**: Article3.

Sobel, E. and K. Lange (1996). "Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics." *Am J Hum Genet* **58**: 1323-37.

Song, J. Y., K. Ichtchenko, T. C. Sudhof and N. Brose (1999). "Neuroligin 1 is a postsynaptic cell-adhesion molecule of excitatory synapses." *Proc Natl Acad Sci U S A* **96**: 1100-5.

St Clair, D., D. Blackwood, W. Muir, A. Carothers, M. Walker, G. Spowart, C. Gosden and H. J. Evans (1990). "Association within a family of a balanced autosomal translocation with major mental illness." *Lancet* **336**: 13-6.

Steffenburg, S., C. Gillberg, L. Hellgren, L. Andersson, I. C. Gillberg, G. Jakobsson and M. Bohman (1989). "A twin study of autism in Denmark, Finland, Iceland, Norway and Sweden." *J Child Psychol Psychiatry* **30**: 405-16.

Stifani, S., C. M. Blaumueller, N. J. Redhead, R. E. Hill and S. Artavanis-Tsakonas (1992). "Human homologs of a Drosophila Enhancer of split gene product define a novel family of nuclear proteins." *Nat Genet* **2**: 119-27.

Stone, J. L., B. Merriman, R. M. Cantor, D. H. Geschwind and S. F. Nelson (2007). "High density SNP association study of a major autism linkage region on chromosome 17." *Hum Mol Genet* **16**: 704-15.

Stone, J. L., B. Merriman, R. M. Cantor, A. L. Yonan, T. C. Gilliam, D. H. Geschwind and S. F. Nelson (2004). "Evidence for sex-specific risk alleles in autism spectrum disorder." *Am J Hum Genet* **75**: 1117-23.

Stranger, B. E., M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavare, P. Deloukas, M. E. Hurles and E. T. Dermitzakis (2007a). "Relative impact of nucleotide and copy number variation on gene expression phenotypes." *Science* **315**: 848-53.

Stranger, B. E., A. C. Nica, M. S. Forrest, A. Dimas, C. P. Bird, C. Beazley, C. E. Ingle, M. Dunning, P. Flicek, D. Koller, S. Montgomery, S. Tavare, P. Deloukas and E. T. Dermitzakis (2007b). "Population genomics of human gene expression." *Nat Genet* **39**: 1217-24.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proc Natl Acad Sci U S A* **102**: 15545-50.

Sudhof, T. C. (2008). "Neuroligins and neurexins link synaptic function to cognitive disease." *Nature* **455**: 903-11.

Sur, M. and J. L. Rubenstein (2005). "Patterning and plasticity of the cerebral cortex." *Science* **310**: 805-10.

Sutcliffe, J. S., R. J. Delahanty, H. C. Prasad, J. L. McCauley, Q. Han, L. Jiang, C. Li, S. E. Folstein and R. D. Blakely (2005). "Allelic heterogeneity at the serotonin transporter

locus (SLC6A4) confers susceptibility to autism and rigid-compulsive behaviors." *Am J Hum Genet* **77**: 265-79.

Szatmari, P., M. B. Jones, L. Zwaigenbaum and J. E. MacLean (1998). "Genetics of autism: overview and new directions." *J Autism Dev Disord* **28**: 351-68.

Szatmari, P., A. D. Paterson, L. Zwaigenbaum, W. Roberts, J. Brian, X. Q. Liu, J. B. Vincent, J. L. Skaug, A. P. Thompson, L. Senman, L. Feuk, C. Qian, S. E. Bryson, M. B. Jones, C. R. Marshall, S. W. Scherer, V. J. Vieland, C. Bartlett, L. V. Mangin, R. Goedken, A. Segre, M. A. Pericak-Vance, M. L. Cuccaro, J. R. Gilbert, H. H. Wright, R. K. Abramson, C. Betancur, T. Bourgeron, C. Gillberg, M. Leboyer, J. D. Buxbaum, K. L. Davis, E. Hollander, J. M. Silverman, J. Hallmayer, L. Lotspeich, J. S. Sutcliffe, J. L. Haines, S. E. Folstein, J. Piven, T. H. Wassink, V. Sheffield, D. H. Geschwind, M. Bucan, W. T. Brown, R. M. Cantor, J. N. Constantino, T. C. Gilliam, M. Herbert, C. Lajonchere, D. H. Ledbetter, C. Lese-Martin, J. Miller, S. Nelson, C. A. Samango-Sprouse, S. Spence, M. State, R. E. Tanzi, H. Coon, G. Dawson, B. Devlin, A. Estes, P. Flodman, L. Klei, W. M. McMahon, N. Minshew, J. Munson, E. Korvatska, P. M. Rodier, G. D. Schellenberg, M. Smith, M. A. Spence, C. Stodgell, P. G. Tepper, E. M. Wijsman, C. E. Yu, B. Roge, C. Mantoulan, K. Wittemeyer, A. Poustka, B. Felder, S. M. Klauck, C. Schuster, F. Poustka, S. Bolte, S. Feineis-Matthews, E. Herbrecht, G. Schmotzer, J. Tsiantis, K. Papanikolaou, E. Maestrini, E. Bacchelli, F. Blasi, S. Carone, C. Toma, H. Van Engeland, M. de Jonge, C. Kemner, F. Koop and M. Langemeijer (2007). "Mapping autism risk loci using genetic linkage and chromosomal rearrangements." *Nat Genet* **39**: 319-28.

Tabuchi, K., J. Blundell, M. R. Etherton, R. E. Hammer, X. Liu, C. M. Powell and T. C. Sudhof (2007). "A neuroligin-3 mutation implicated in autism increases inhibitory synaptic transmission in mice." *Science* **318**: 71-6.

Talebizadeh, Z., M. G. Butler and M. F. Theodoro (2008). "Feasibility and relevance of examining lymphoblastoid cell lines to study role of microRNAs in autism." *Autism Res* **1**: 240-50.

Tan, Z., G. Randall, J. Fan, B. Camoretti-Mercado, R. Brockman-Schneider, L. Pan, J. Solway, J. E. Gern, R. F. Lemanske, D. Nicolae and C. Ober (2007). "Allele-specific targeting of microRNAs to HLA-G and risk of asthma." *Am J Hum Genet* **81**: 829-34.

Tapper, W. J., N. Maniatis, N. E. Morton and A. Collins (2003). "A metric linkage disequilibrium map of a human chromosome." *Ann Hum Genet* **67**: 487-94.

Tarpey, P. S., R. Smith, E. Pleasance, A. Whibley, S. Edkins, C. Hardy, S. O'Meara, C. Latimer, E. Dicks, A. Menzies, P. Stephens, M. Blow, C. Greenman, Y. Xue, C. Tyler-Smith, D. Thompson, K. Gray, J. Andrews, S. Barthorpe, G. Buck, J. Cole, R. Dunmore, D. Jones, M. Maddison, T. Mironenko, R. Turner, K. Turrell, J. Varian, S. West, S. Widaa, P. Wray, J. Teague, A. Butler, A. Jenkinson, M. Jia, D. Richardson, R. Shepherd, R. Wooster, M. I. Tejada, F. Martinez, G. Carvill, R. Goliath, A. P. de Brouwer, H. van Bokhoven, H. Van Esch, J. Chelly, M. Raynaud, H. H. Ropers, F. E. Abidi, A. K. Srivastava, J. Cox, Y. Luo, U. Mallya, J. Moon, J. Parnau, S. Mohammed, J. L. Tolmie, C. Shoubridge, M. Corbett, A. Gardner, E. Haan, S. Rujirabanjerd, M. Shaw, L. Vandeleur, T. Fullston, D. F. Easton, J. Boyle, M. Partington, A. Hackett, M. Field, C. Skinner, R. E. Stevenson, M. Bobrow, G. Turner, C. E. Schwartz, J. Gecz, F. L. Raymond, P. A. Futreal and M. R. Stratton (2009). "A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation." *Nat Genet* **41**: 535-43.

Terwilliger, J. and J. Ott (1994). Handbook of human genetic linkage. Baltimore, The John Hopkins University Press.

Teslovich, T. M., K. Musunuru, A. V. Smith, A. C. Edmondson, I. M. Stylianou, M. Koseki, J. P. Pirruccello, S. Ripatti, D. I. Chasman, C. J. Willer, C. T. Johansen, S. W. Fouchier, A.

Isaacs, G. M. Peloso, M. Barbalic, S. L. Ricketts, J. C. Bis, Y. S. Aulchenko, G. Thorleifsson, M. F. Feitosa, J. Chambers, M. Orho-Melander, O. Melander, T. Johnson, X. Li, X. Guo, M. Li, Y. Shin Cho, M. Jin Go, Y. Jin Kim, J. Y. Lee, T. Park, K. Kim, X. Sim, R. Twee-Hee Ong, D. C. Croteau-Chonka, L. A. Lange, J. D. Smith, K. Song, J. Hua Zhao, X. Yuan, J. Luan, C. Lamina, A. Ziegler, W. Zhang, R. Y. Zee, A. F. Wright, J. C. Witteman, J. F. Wilson, G. Willemsen, H. E. Wichmann, J. B. Whitfield, D. M. Waterworth, N. J. Wareham, G. Waeber, P. Vollenweider, B. F. Voight, V. Vitart, A. G. Uitterlinden, M. Uda, J. Tuomilehto, J. R. Thompson, T. Tanaka, I. Surakka, H. M. Stringham, T. D. Spector, N. Soranzo, J. H. Smit, J. Sinisalo, K. Silander, E. J. Sijbrands, A. Scuteri, J. Scott, D. Schlessinger, S. Sanna, V. Salomaa, J. Saharinen, C. Sabatti, A. Ruokonen, I. Rudan, L. M. Rose, R. Roberts, M. Rieder, B. M. Psaty, P. P. Pramstaller, I. Pichler, M. Perola, B. W. Penninx, N. L. Pedersen, C. Pattaro, A. N. Parker, G. Pare, B. A. Oostra, C. J. O'Donnell, M. S. Nieminen, D. A. Nickerson, G. W. Montgomery, T. Meitinger, R. McPherson and M. I. McCarthy (2010). "Biological, clinical and population relevance of 95 loci for blood lipids." *Nature* **466**: 707-13.

The International HapMap Consortium (2003). "The International HapMap Project." *Nature* **426**: 789-96.

The International HapMap Consortium (2005). "A haplotype map of the human genome." *Nature* **437**: 1299-320.

Thomas, N. S., A. J. Sharp, C. E. Browne, D. Skuse, C. Hardie and N. R. Dennis (1999). "Xp deletions associated with autism in three females." *Hum Genet* **104**: 43-8.

Thomson, P. A., S. E. Harris, J. M. Starr, L. J. Whalley, D. J. Porteous and I. J. Deary (2005a). "Association between genotype at an exonic SNP in DISC1 and normal cognitive aging." *Neurosci Lett* **389**: 41-5.

Thomson, P. A., N. R. Wray, J. K. Millar, K. L. Evans, S. L. Hellard, A. Condie, W. J. Muir, D. H. Blackwood and D. J. Porteous (2005b). "Association between the TRAX/DISC locus and both bipolar disorder and schizophrenia in the Scottish population." *Mol Psychiatry* **10**: 657-68, 616.

Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J. M. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, M. W. Smith, M. A. Thera, C. Wambebe, J. L. Weber and S. M. Williams (2009). "The genetic structure and history of Africans and African Americans." *Science* **324**: 1035-44.

Tomppo, L., W. Hennah, P. Lahermo, A. Loukola, A. Tuulio-Henriksson, J. Suvisaari, T. Partonen, J. Ekelund, J. Lonnqvist and L. Peltonen (2009). "Association between genes of Disrupted in schizophrenia 1 (DISC1) interactors and schizophrenia supports the role of the DISC1 pathway in the etiology of major mental illnesses." *Biol Psychiatry* **65**: 1055-62.

Trikalinos, T. A., A. Karvouni, E. Zintzaras, T. Ylisaukko-oja, L. Peltonen, I. Jarvela and J. P. Ioannidis (2006). "A heterogeneity-based genome search meta-analysis for autism-spectrum disorders." *Mol Psychiatry* **11**: 29-36.

Uda, M., R. Galanello, S. Sanna, G. Lettre, V. G. Sankaran, W. Chen, G. Usala, F. Busonero, A. Maschio, G. Albai, M. G. Piras, N. Sestu, S. Lai, M. Dei, A. Mulas, L. Crisponi, S. Naitza, I. Asunis, M. Deiana, R. Nagaraja, L. Perseu, S. Satta, M. D. Cipollina, C. Sollaino, P. Moi, J. N. Hirschhorn, S. H. Orkin, G. R. Abecasis, D. Schlessinger and A. Cao (2008). "Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia." *Proc Natl Acad Sci U S A* **105**: 1620-5.

Vandenplas, S., I. Wiid, A. Grobler-Rabie, K. Brebner, M. Ricketts, G. Wallis, A. Bester, C. Boyd and C. Mathew (1984). "Blot hybridisation analysis of genomic DNA." *J Med Genet* **21**: 164-72.

Varilo, T. (1999). The age of the mutations in the Finnish disease heritage; a genealogical and linkage disequilibrium study. Helsinki, National Public Health Institute and University of Helsinki.

Varilo, T., K. Nikali, A. Suomalainen, T. Lonnqvist and L. Peltonen (1996). "Tracing an ancestral mutation: genealogical and haplotype analysis of the infantile onset spinocerebellar ataxia locus." *Genome Res* **6**: 870-5.

Varilo, T., T. Paunio, A. Parker, M. Perola, J. Meyer, J. D. Terwilliger and L. Peltonen (2003). "The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories." *Hum Mol Genet* **12**: 51-9.

Varoqueaux, F., S. Jamain and N. Brose (2004). "Neuroligin 2 is exclusively localized to inhibitory synapses." *Eur J Cell Biol* **83**: 449-56.

Vasudevan, S. and J. A. Steitz (2007). "AU-rich-element-mediated upregulation of translation by FXR1 and Argonaute 2." *Cell* **128**: 1105-18.

Veenstra-VanderWeele, J. and E. H. Cook, Jr. (2004). "Molecular genetics of autism spectrum disorder." *Mol Psychiatry* **9**: 819-32.

Vionnet, N., M. Stoffel, J. Takeda, K. Yasuda, G. I. Bell, H. Zouali, S. Lesage, G. Velho, F. Iris, P. Passa and et al. (1992). "Nonsense mutation in the glucokinase gene causes early-onset non-insulin-dependent diabetes mellitus." *Nature* **356**: 721-2.

Visscher, P. M. (2008). "Sizing up human height variation." *Nat Genet* **40**: 489-90.

Volkmar, F. and A. Klin (2000). Diagnostic issues in Asperger syndrome. Asperger Syndrome. A. Klin, F. Volkmar and S. Sparrow. New York, US, The Guilford Press**:** 25-71.

Volkmar, F. R., D. J. Cohen, J. D. Bregman, M. Y. Hooks and J. M. Stevenson (1989). "An examination of social typologies in autism." *J Am Acad Child Adolesc Psychiatry* **28**: 82-6.

Volkmar, F. R., A. Klin and D. Pauls (1998). "Nosological and genetic aspects of Asperger syndrome." *J Autism Dev Disord* **28**: 457-63.

Walsh, C. A., E. M. Morrow and J. L. Rubenstein (2008). "Autism and brain development." *Cell* **135**: 396-400.

Wang, K., M. Li and M. Bucan (2007). "Pathway-Based Approaches for Analysis of Genomewide Association Studies." *Am J Hum Genet* **81**.

Wang, K., H. Zhang, S. Kugathasan, V. Annese, J. P. Bradfield, R. K. Russell, P. M. Sleiman, M. Imielinski, J. Glessner, C. Hou, D. C. Wilson, T. Walters, C. Kim, E. C. Frackelton, P. Lionetti, A. Barabino, J. Van Limbergen, S. Guthery, L. Denson, D. Piccoli, M. Li, M. Dubinsky, M. Silverberg, A. Griffiths, S. F. Grant, J. Satsangi, R. Baldassano and H. Hakonarson (2009a). "Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease." *Am J Hum Genet* **84**: 399-405.

Wang, K., H. Zhang, D. Ma, M. Bucan, J. T. Glessner, B. S. Abrahams, D. Salyakina, M. Imielinski, J. P. Bradfield, P. M. Sleiman, C. E. Kim, C. Hou, E. Frackelton, R. Chiavacci, N. Takahashi, T. Sakurai, E. Rappaport, C. M. Lajonchere, J. Munson, A. Estes, O. Korvatska, J. Piven, L. I. Sonnenblick, A. I. Alvarez Retuerto, E. I. Herman, H. Dong, T. Hutman, M. Sigman, S. Ozonoff, A. Klin, T. Owley, J. A. Sweeney, C. W. Brune, R. M. Cantor, R. Bernier, J. R. Gilbert, M. L. Cuccaro, W. M. McMahon, J. Miller, M. W. State, T. H. Wassink, H. Coon, S. E. Levy, R. T. Schultz, J. I. Nurnberger, J. L. Haines, J. S. Sutcliffe, E. H. Cook, N. J. Minshew, J. D. Buxbaum, G. Dawson, S. F. Grant, D. H. Geschwind, M. A. Pericak-Vance, G. D. Schellenberg and H.

Hakonarson (2009b). "Common genetic variants on 5p14.1 associate with autism spectrum disorders." *Nature* **459**: 528-33.

Wang, W. X., B. W. Rajeev, A. J. Stromberg, N. Ren, G. Tang, Q. Huang, I. Rigoutsos and P. T. Nelson (2008). "The expression of microRNA miR-107 decreases early in Alzheimer's disease and may accelerate disease progression through regulation of beta-site amyloid precursor protein-cleaving enzyme 1." *J Neurosci* **28**: 1213-23.

Wassink, T. H., J. Piven and S. R. Patil (2001a). "Chromosomal abnormalities in a clinic sample of individuals with autistic disorder." *Psychiatr Genet* **11**: 57-63.

Wassink, T. H., J. Piven, V. J. Vieland, J. Huang, R. E. Swiderski, J. Pietila, T. Braun, G. Beck, S. E. Folstein, J. L. Haines and V. C. Sheffield (2001b). "Evidence supporting WNT2 as an autism susceptibility gene." *Am J Med Genet* **105**: 406-13.

Weber, J. L. and C. Wong (1993). "Mutation of human short tandem repeats." *Hum Mol Genet* **2**: 1123-8.

Weiss, L. A., D. E. Arking, M. J. Daly and A. Chakravarti (2009). "A genome-wide linkage and association scan reveals novel loci for autism." *Nature* **461**: 802-8.

Weiss, L. A., Y. Shen, J. M. Korn, D. E. Arking, D. T. Miller, R. Fossdal, E. Saemundsen, H. Stefansson, M. A. Ferreira, T. Green, O. S. Platt, D. M. Ruderfer, C. A. Walsh, D. Altshuler, A. Chakravarti, R. E. Tanzi, K. Stefansson, S. L. Santangelo, J. F. Gusella, P. Sklar, B. L. Wu and M. J. Daly (2008). "Association between Microdeletion and Microduplication at 16p11.2 and Autism." *N Engl J Med.*

Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature* **447**: 661-78.

Willemsen, M. H., B. A. Fernandez, C. A. Bacino, E. Gerkes, A. P. de Brouwer, R. Pfundt, B. Sikkema-Raddatz, S. W. Scherer, C. R. Marshall, L. Potocki, H. van Bokhoven and T. Kleefstra (2010). "Identification of ANKRD11 and ZNF778 as candidate genes for autism and variable cognitive impairment in the novel 16q24.3 microdeletion syndrome." *Eur J Hum Genet* **18**: 429-35.

Williams, J. M., T. F. Beck, D. M. Pearson, M. B. Proud, S. W. Cheung and D. A. Scott (2009). "A 1q42 deletion involving DISC1, DISC2, and TSNAX in an autism spectrum disorder." *Am J Med Genet A* **149A**: 1758-62.

Wing, L. and J. Gould (1979). "Severe impairments of social interaction and associated abnormalities in children: epidemiology and classification." *J Autism Dev Disord* **9**: 11-29.

World Health Organization (1993). The ICD-10 Classification of Mental and Behavioural Disorders. Diagnostic Criteria for Research. Geneva, WHO.

Xie, X., J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander and M. Kellis (2005). "Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals." *Nature* **434**: 338-45.

Xu, S., X. Yin, S. Li, W. Jin, H. Lou, L. Yang, X. Gong, H. Wang, Y. Shen, X. Pan, Y. He, Y. Yang, Y. Wang, W. Fu, Y. An, J. Wang, J. Tan, J. Qian, X. Chen, X. Zhang, Y. Sun, X. Zhang, B. Wu and L. Jin (2009). "Genomic dissection of population substructure of Han Chinese and its implication in association studies." *Am J Hum Genet* **85**: 762-74.

Yamada, K., K. Nakamura, Y. Minabe, Y. Iwayama-Shigeno, H. Takao, T. Toyota, E. Hattori, N. Takei, Y. Sekine, K. Suzuki, Y. Iwata, K. Miyoshi, A. Honda, K. Baba, T. Katayama, M. Tohyama, N. Mori and T. Yoshikawa (2004). "Association analysis of FEZ1 variants with schizophrenia in Japanese cohorts." *Biol Psychiatry* **56**: 683-90.

Yan, J., G. Oliveira, A. Coutinho, C. Yang, J. Feng, C. Katz, J. Sram, A. Bockholt, I. R. Jones, N. Craddock, E. H. Cook, Jr., A. Vicente and S. S. Sommer (2005). "Analysis of the neuroligin 3 and 4 genes in autism and other neuropsychiatric patients." *Mol Psychiatry* **10**: 329-32.

Yang, M. S. and M. Gill (2007). "A review of gene linkage, association and expression studies in autism and an assessment of convergent evidence." *Int J Dev Neurosci* **25**: 69-85.

Ylisaukko-oja, T., T. Nieminen-von Wendt, E. Kempas, S. Sarenius, T. Varilo, L. von Wendt, L. Peltonen and I. Jarvela (2004). "Genome-wide scan for loci of Asperger syndrome." *Mol Psychiatry* **9**: 161-8.

Yonan, A. L., M. Alarcon, R. Cheng, P. K. Magnusson, S. J. Spence, A. A. Palmer, A. Grunn, S. H. Juo, J. D. Terwilliger, J. Liu, R. M. Cantor, D. H. Geschwind and T. C. Gilliam (2003). "A genomewide screen of 345 families for autism-susceptibility loci." *Am J Hum Genet* **73**: 886-97.

Yu, A., C. Zhao, Y. Fan, W. Jang, A. J. Mungall, P. Deloukas, A. Olsen, N. A. Doggett, N. Ghebranious, K. W. Broman and J. L. Weber (2001). "Comparison of human genetic and sequence-based physical maps." *Nature* **409**: 951-3.

Zeng, Y., R. Yi and B. R. Cullen (2003). "MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms." *Proc Natl Acad Sci U S A* **100**: 9779-84.

Zhang, F., J. Sarginson, C. Crombie, N. Walker, D. St Clair and D. Shaw (2006). "Genetic association between schizophrenia and the DISC1 gene in the Scottish population." *Am J Med Genet B Neuropsychiatr Genet* **141**: 155-9.

Zhang, J., R. P. Finney, R. J. Clifford, L. K. Derr and K. H. Buetow (2005). "Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach." *Genomics* **85**: 297-308.

Zhong, H., X. Yang, L. M. Kaplan, C. Molony and E. E. Schadt (2010). "Integrating pathway analysis and genetics of gene expression for genome-wide association studies." *Am J Hum Genet* **86**: 581-91.